Cakrawala Pendidikan
Jurnal Ilmiah Pendidikan
Vol. 45 No. 2, June 2026, pp.340-350
https://journal.uny.ac.id/index.php/cp/issue/view/3110
DOI: https://doi.org/10.21831/cp.v45i2.94152

Assessing motor competence in Indonesian sports science admissions: A comparison of portfolio assessment and direct practical testing

Ngadiman[1]*, Akbar Kusuma Abadi[2], Eko Sumianto[1], Agung Prabowo[1], Bayu Suko Wahono[1]
[1]Universitas Jenderal Soedirman, Indonesia
[2]Universitas Negeri Semarang, Indonesia
*Corresponding Author: ngadiman@unsoed.ac.id

ABSTRACT

Indonesia's National Selection for State University Admission (SNMPTN) for Sports Science programs has shifted from practical motor skill tests to portfolio-based assessment, raising concerns about validity and reliability in measuring motor competence. This study examined the validity and reliability of portfolio assessment by comparing it with direct motor skill tests as the criterion measure. A quantitative descriptive-verificative design involved 50 Physical Education students at Jenderal Soedirman University with complete portfolio and practical test data. Four motor skill domains were assessed: hand-eye coordination, agility, lower-limb muscle power, and endurance. Statistical analyses included Wilcoxon test, Spearman's correlation, Intraclass Correlation Coefficient (ICC), Bland-Altman analysis, and error measurements. Results showed portfolio scores consistently overestimated performance with significant differences in three domains ($p \leq 0.001$). Spearman's correlations ranged from moderate to strong ($\rho = 0.432–0.814$), yet ICC values indicated low-to-moderate absolute agreement except for lower-limb muscle power (ICC = 0.800). Bland-Altman analysis revealed systematic positive bias with wide limits of agreement. These findings suggest portfolio assessment is inadequate as a standalone instrument for selecting candidates based on motor skills and should serve as a complementary tool to standardize direct practical tests in Sports Science admissions.

Keywords: assessment, portfolio, sports, science,admissions

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
25 December 2025	*27 January 2026*	*18 March 2026*	*27 May 2026*

Citation (APA Style): Ngadiman, N., Abadi, A. K., Sumianto, E., Prabowo, A., & Wahono, B. S. (2026). Assessing motor competence in Indonesian sports science admissions: A comparison of portfolio assessment and direct practical testing. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 45*(2), pp.340-350. DOI: https://doi.org/10.21831/cp.v45i2.94152

INTRODUCTION

Since 2019, the change in the national selection policy for prospective students of sports science programs from direct practice-based motor skill tests to portfolio-based assessment has represented a significant shift in the assessment paradigm. This shift carries epistemological implications, as motor competence is performative and, from a theoretical standpoint, is better measured through direct observation than through indirect documentation (Yang et al., 2025). Consequently, the basis for assessing the physical and motor readiness of prospective students has also been reoriented (Choo et al., 2024).

In the direct practical test model, assessors can observe participants' physical performance in real time. Various aspects, such as strength, agility, coordination, balance, technical quality, responsiveness to instructions, rhythm, and movement consistency, can be comprehensively evaluated under relatively uniform conditions (Abudari et al., 2022; Makaruk et al., 2023). This information serves as an important indicator of physiological readiness and baseline motor skills that are relevant to the academic demands in the field of sports science (Eddy et al., 2020).

This approach aligns with the principles of authentic assessment in physical education. In physical education, the purpose of authentic assessment is to capture the cognitive, affective, and psychomotor dimensions of student learning. However, the implementation of authentic assessment is not an easy endeavor, particularly when teachers are required to assess students' movement skills using detailed and consistent assessment criteria (Hariadi et al., 2025). This indicates that assessment in physical education cannot rely solely on documentation but still requires a valid standard procedure that is capable of representing actual motor performance.

Conversely, the portfolio system shifts the assessment process from the performance arena to the analysis of documents, such as photos, videos, certificates, and activity reports. This orientation makes the selection outcomes more dependent on the ability to compile evidence than on actual performance (de Jong et al., 2022; Pool et al., 2020). In addition, the time gap between portfolio creation and the implementation of the selection process introduces the potential for time-lag bias, as changes in physical condition, fitness level, and injury status are not always reflected in static documents.

Institutional contextual factors further undermine the fairness of the assessment. Disparities in school facilities make the quality of portfolio documentation heavily dependent on the availability of recording resources. Schools with adequate facilities are better able to produce portfolios that appear professional, regardless of whether the participants' motor abilities are actually higher (Hogan et al., 2023; Penney et al., 2023). Conversely, variations in the competence of sports teachers or mentors affect the structure, completeness, and relevance of portfolio content, thereby increasing the heterogeneity of the quality of evidence being assessed (van Maarseveen et al., 2025).

The technical quality of portfolios also often fails to meet the prerequisites for accurate movement observation. Errors in camera angle, disproportionate camera distance, low visual quality, poor lighting, and documentation that does not display the full sequence of movements limit assessors' ability to observe technique and coordination (Liu & Li, 2022; Ng & Button, 2023). These conditions undermine construct validity, as the assessment indicators are not adequately reflected in the available evidence.

Fairness in assessment in physical education is also influenced by the objectivity of the assessor. Previous research in physical education courses at the university level found that assessment results vary considerably. This variation depends on the characteristics of the lecturer conducting the assessment, including the lecturer's teaching experience, academic background, teaching load, and the assessment standards used (Kristiyandaru et al., 2023). Therefore, when portfolio evidence is used in high-stakes selection, differences in assessor interpretation may threaten the reliability and fairness of the selection process.

From a psychometric perspective, the reliability of portfolio assessment becomes problematic. The same document may yield different scores because it is highly dependent on the assessor's individual ability and interpretation of the visualized movements, whereas direct practical tests provide a more uniform observation context (Carballo-Fazanes et al., 2021; Ross et al., 2024). In addition, the opportunity to select the best moments, re-record videos multiple times, and adjust camera angles to conceal technical weaknesses increases the risk of discrepancies between documented performance and actual motor ability (Hulteen et al., 2023; Rey et al., 2020).

Empirically, several sports science study programs have identified discrepancies between performance that appears superior in portfolios and students' actual motor abilities after admission, particularly in basic practical courses. This condition results in reduced initial learning efficiency and an increased need for remedial programs. These field observations are consistent with previous research indicating that direct tests provide higher assessment accuracy and more stable inter-rater reliability, and that portfolios are more appropriately positioned as supporting assessments rather than as the primary measure of motor performance.

Despite growing concerns about portfolio-based assessment in sports science admissions globally, empirical research specifically examining the psychometric properties of portfolio assessment as a replacement for direct motor skill testing remains limited. While previous studies have examined portfolio validity in academic settings or talent identification contexts (Penney et

al., 2023), no studies have systematically compared portfolio assessment with direct testing using no studies have systematically compared portfolio assessment with direct testing using comprehensive psychometric approaches, including the Wilcoxon signed-rank test, Spearman correlation, ICC, Bland-Altman analysis, and error metrics specifically in the context of national university admissions for sports science programs. The Indonesian context presents a unique case where this policy shift affects thousands of candidates annually through the national selection system (SNMPTN), yet no empirical validation has been conducted to assess whether this substitution maintains measurement validity and fairness. This study addresses this gap by providing the first comprehensive psychometric evaluation of portfolio assessment validity in Indonesian sports science admissions, with direct implications for evidence-based policy reform in selection systems that have similarly transitioned from direct to indirect assessment methods.

Given these issues of validity, reliability, and assessment fairness, the need for a scientific evaluation of the portfolio-based selection system has become urgent. This study was designed to examine the validity and reliability of portfolio assessment by comparing it with direct motor skill tests administered by lecturers. The findings are expected to provide an empirical basis for improving selection policies, making assessment mechanisms for prospective sports science students more objective, fair, and academically defensible.

METHOD

This study employed a quantitative approach to evaluate the validity and reliability of portfolio assessment data in relation to the motor skills of students in sports science study programs who had been admitted through the national selection pathway. The quantitative approach was chosen because it produces objective data that can be systematically analyzed through numerical procedures and allows for testing the degree of concordance between two forms of assessment. The characteristics of this approach are aligned with the aim of the study, which focuses on measuring the relationship, consistency, and accuracy of the two assessment methods used in the admission process.

This study employed a criterion-related validation design to examine the concurrent validity of portfolio assessment by comparing it against direct motor skill testing as the criterion measure. The descriptive component provides an overview of the distribution of portfolio scores and motor skill test scores, whereas the validation component examines the validity and reliability of the two methods. The analyses included the Wilcoxon signed-rank test to assess the agreement between measurement results, the Intraclass Correlation Coefficient (ICC) to evaluate absolute consistency, the Bland Altman analysis to examine systematic bias between methods, mixed-effects models to identify sources of score variation, and Mean Squared Error (MSE) and Mean Absolute Error (MAE) to quantify the accuracy of portfolio scores relative to direct tests. This design was selected to determine whether portfolio assessment demonstrates sufficient psychometric properties to function as a valid substitute for or complement to direct practical testing in high-stakes selection contexts.

The study population consisted of all students in the Physical Education Study Program at Universitas Jenderal Soedirman, academic year 2025/2026, who were admitted through the national selection pathway (SNMPTN) and had submitted portfolio assessment documents, while direct motor skill testing was conducted by university lecturers during the first semester. The sample consisted of 50 students from the Physical Education Study Program at Jenderal Soedirman University who had complete data for both portfolio assessments and direct practical tests. A purposive sampling technique was used because only participants with complete scores for both types of assessment were included in the comparative analysis.

The research data was obtained from two forms of measurement. First, the portfolio assessment scores were derived from practice tasks performed independently by the participants, who submitted recordings of physical activities and evidence of motor skills in accordance with the national selection format. Second, the motor skill test scores were obtained from direct practical examinations conducted under the supervision of academic assessors to capture actual physical performance. Both types of measurement used the same skill indicators; however, their

implementation differed, as portfolio tasks were self-administered by participants, whereas direct tests were conducted face-to-face by lecturers.

Data collection began with checking the completeness of the portfolio documents, obtaining motor skill test scores from the study program, and coding both types of data into an analytical format. Score entry was carried out using standardized procedures to ensure comparability of scales across methods. Consistency checks of the indicators were performed to ensure that the portfolio assessment and motor tests applied equivalent assessment criteria. The database was compiled through a multi-stage verification process to prevent entry errors prior to statistical analysis.

The data were analyzed using statistical software to examine the relationship, agreement, and accuracy of portfolio scores relative to direct test results. Regression analysis was used to evaluate predictive validity. The Intraclass Correlation Coefficient was employed to assess the reliability of consistency between methods. Bland–Altman analysis was applied to detect systematic differences and to determine the limits of agreement between the two methods. Mixed-effects models were used to identify sources of score deviation attributable to the assessment method, individual differences, or technical variation. Mean Squared Error and Mean Absolute Error were calculated to quantify the accuracy of portfolio scores compared with direct test scores. The analytical results were used to determine the feasibility of using portfolios as a valid and reliable selection method.

FINDINGS AND DISCUSSION

Findings

Description of portfolio and direct test scores

Descriptive statistics were used to provide an overview of portfolio assessment scores and direct test scores for each measured skill domain. The mean and standard deviation (SD) for the four domains, hand-eye coordination, agility, lower-limb muscle power, and endurance, obtained from both portfolio assessment and direct testing are presented in Table 1.

Table 1. Descriptive statistics of portfolio and direct test scores

No.	Domain	N	Porto Mean ± SD	Direct Mean ± SD
1.	Hand–eye coordination	50	28.88 ± 5.84	22.92 ± 5.86
2.	Agility	50	17.50 ± 1.79	18.68 ± 1.61
3.	Lower-limb muscle power	50	52.76 ± 11.69	50.72 ± 11.73
4.	Endurance	50	7.84 ± 1.92	8.33 ± 1.31

Descriptive analysis shows that portfolio scores consistently indicated higher performance than direct test scores across all domains. For hand-eye coordination, portfolio scores (M = 28.88, SD = 5.84) exceeded direct test scores (M = 22.92, SD = 5.86) by 5.96 points (26% difference). For lower-limb muscle power, portfolio scores (M = 52.76, SD = 11.69) were 2.04 points higher than direct tests (M = 50.72, SD = 11.73), representing a 4% difference. For time-based measures (where lower scores indicate better performance), portfolio scores showed shorter completion times: agility portfolio (M = 17.50 s, SD = 1.79) versus direct test (M = 18.68 s, SD = 1.61), a difference of 1.18 seconds (6.3% faster); endurance portfolio (M = 7.84 m, SD = 1.92) versus direct test (M = 8.33 m, SD = 1.31), a difference of 5.9 minutes (5.9% faster).

Comparison of portfolio and direct test scores (Wilcoxon test)

This analysis examined whether the observed descriptive differences between portfolio assessment and direct testing scores in the same participants (n = 50) were statistically significant. Before selecting an appropriate paired-difference test, normality was assessed for each score distribution using the Shapiro-Wilk test. At $\alpha = 0.05$, data were considered normally distributed when $p > 0.05$.

The Shapiro-Wilk results (see Table 2) indicate that the normality assumption was not consistently satisfied across domains. Given these findings, paired comparisons between portfolio

and direct test scores were conducted using the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test.

Table 2. Shapiro–wilk normality test results (n = 50)

No.	Variable	W	df	p-value	Decision
1.	Portfolio Hand–eye coordination	0.921	50	0.002	Not normal
2.	Portfolio Agility	0.950	50	0.035	Not normal
3.	Portfolio Lower-limb muscle power	0.970	50	0.231	Normal
4.	Portfolio Endurance	0.802	50	< 0.001	Not normal
5.	Direct Hand–eye coordination	0.973	50	0.294	Normal
6.	Direct Agility	0.941	50	0.015	Not normal
7.	Direct Lower-limb muscle power	0.954	50	0.052	Normal
8.	Direct Endurance	0.931	50	0.006	Not normal

Table 3. Wilcoxon test results for portfolio vs direct test scores

No.	Domain	N	Mean Porto	Mean Direct	Z	p (2-tailed)
1.	Hand–eye coordination	50	28.88	22.92	−5.069	< 0.001
2.	Agility	50	17.50	18.68	−4.407	< 0.001
3.	Lower-limb muscle power	50	52.76	50.72	−1.829	0.067
4.	Endurance	50	7.84	8.33	−3.186	0.001

The Wilcoxon signed-rank test results (see Table 3) reveal statistically significant differences between portfolio and direct test scores for hand-eye coordination ($Z = -5.069$, $p < 0.001$), agility ($Z = -4.407$, $p < 0.001$), and endurance ($Z = -3.186$, $p = 0.001$), whereas the difference in lower-limb muscle power approached but did not reach statistical significance ($Z = -1.829$, $p = 0.067$).

Convergent validity: Spearman correlation

Convergent validity was analyzed using Spearman's rank-order correlation coefficient to examine the degree to which portfolio and direct test scores rank-order participants similarly (Table 4).

Table 4. Spearman correlations between portfolio and direct tests

No.	Domain	ρ Spearman	p (2-tailed)	Category
1.	Hand–eye coordination	0.432	0.002	Moderate
2.	Agility	0.598	< 0.001	Moderate approaching strong
3.	Lower-limb muscle power	0.814	< 0.001	Very strong
4.	Endurance	0.624	< 0.001	Strong

The Spearman correlations between portfolio and direct test scores were positive and statistically significant across all four domains ($p < 0.01$), indicating that both methods tend to rank order participants similarly. Correlation strength varied substantially: hand-eye coordination showed a moderate association ($\rho = 0.432$), agility and endurance showed moderate to strong associations ($\rho = 0.598$ and 0.624, respectively), while lower-limb muscle power exhibited a very strong association ($\rho = 0.814$).

Intraclass Correlation Coefficient (ICC) reliability

While Spearman correlation assesses relative consistency in rank ordering, absolute agreement reliability between portfolio assessments and direct tests was analyzed using a two-way mixed-effects absolute-agreement single-measure Intraclass Correlation Coefficient (ICC) model (Table 5).

The ICC values indicate that absolute agreement between portfolio scores and direct test scores varies considerably across domains. Hand-eye coordination showed low agreement (ICC = 0.237), agility and endurance showed moderate agreement (ICC = 0.417 and 0.428, respectively), while only lower-limb muscle power demonstrated good agreement (ICC = 0.800).

Table 5. ICC Values between Portfolio Assessment and Direct Tests

Domain	ICC (single)	95% CI	Category
Hand–eye coordination	0.237	-0.054 – 0.494	Low
Agility (seconds)	0.417	0.077 – 0.651	Moderate
Lower-limb muscle power	0.800	0.669 – 0.882	Good
Endurance (minutes)	0.428	0.179 – 0.627	Moderate

Bias and limits of agreement analysis: Bland–Altman

A Bland-Altman analysis was conducted to assess the magnitude and direction of measurement differences and to establish the limits of agreement (LoA) between the two methods (Table 6).

Table 6. Summary of Bland–Altman Analysis

Domain	Bias	SD	95% CI bias	LoA		Bias description
				Lower	Upper	
Hand–eye coordination	5.96	6.63	4.07 s.d. 7.85	−7.04	18.96	Positive bias, significant
Agility	1.18	1.68	0.70 s.d. 1.66	−2.12	4.48	Positive bias, significant
Lower-limb muscle power	2.04	7.23	−0.02 s.d. 4.10	−12.14	16.22	Small, non-significant bias
Endurance	0.49	1.74	0.00 s.d. 0.99	−2.91	3.90	Positive, marginal bias

Note: *LoA (limits of agreement) = mean difference ± 1.96 × SD of the differences.

The Bland-Altman analysis reveals varying bias patterns across domains. Hand-eye coordination and agility show consistent positive bias (portfolio scores higher), with mean biases of 5.96 points (95% CI: 4.07 to 7.85) and 1.18 seconds (95% CI: 0.70 to 1.66), respectively. Lower-limb muscle power shows a small, non-significant bias (2.04 points; 95% CI: −0.02 to 4.10). Endurance demonstrates marginal positive bias (0.49 minutes; 95% CI: 0.00 to 0.99). However, all domains display extremely wide limits of agreement hand-eye coordination (−7.04 to 18.96 points; 26-point range), agility (−2.12 to 4.48 seconds; 6.6-second range), lower-limb muscle power (−12.14 to 16.22 points; 28-point range), and endurance (−2.91 to 3.90 minutes; 6.8-minute range), indicating substantial individual-level variability that makes individual predictions highly unreliable for all domains.

Quantitative accuracy: MAE and MSE

The accuracy of portfolio assessment relative to direct tests was analyzed using Mean Absolute Error (MAE) and Mean Squared Error (MSE), with Root Mean Squared Error (RMSE) derived from MSE (Table 7).

Table 7. MAE, MSE, and RMSE Values for Portfolio Assessment

Domain	Definition of difference	MAE	MSE	RMSE
Hand–eye coordination	Portfolio − Direct (score)	7.32 points	78.64	8.87
Agility (seconds)	Direct − Portfolio (time)	1.59 seconds	4.17	2.04
Lower-limb muscle power	Portfolio − Direct (score)	5.56 points	55.44	7.45
Endurance (minutes)	Direct − Portfolio (time)	1.22 minutes	3.21	1.79

The MAE values show that portfolio scores differ on average by 7.32 points for hand-eye coordination (32% of direct test mean), 1.59 seconds for agility (8.5% of direct test mean), 5.56 points for lower-limb power (11% of direct test mean), and 1.22 minutes for endurance (15 % of direct test mean). The RMSE values are consistently larger than the MAE values across all domains.

Mixed-effects model analysis

Mixed-effects models were fitted for each domain to decompose variance and identify sources of variation attributable to measurement method, individual differences, and residual error (Table 8).

Table 8. Key parameters of the mixed-effects models

Domain	Intercept		F(1.99)	p	Residual variance (σ^2)	Residual SD
	Mean	SE				
Hand–eye coordination	25.90	0.65	1565.17	< 0.001	42.86	6.55
Agility (seconds)	18.09	0.18	10132.86	< 0.001	3.23	1.80
Lower-limb muscle power	51.74	1.17	1957.18	< 0.001	136.78	11.70
Endurance (minutes)	8.21	0.32	658.78	< 0.001	10.13	3.18

The fixed-effect tests demonstrated that all intercepts differed significantly from zero (all $p < 0.001$). Residual variance components revealed domain-specific heterogeneity: agility showed the lowest variance ($\sigma^2 = 3.23$), while lower-limb muscle power ($\sigma^2 = 136.78$) and endurance ($\sigma^2 = 10.13$ minutes^2) showed substantially higher variance, indicating high individual-level variability in portfolio-direct test agreement.

Discussion

The results of the descriptive analysis indicate that portfolio assessment scores consistently exceed direct test scores across all measured domains. In the hand–eye coordination domain, the difference reached 26%, whereas in lower-limb muscle power it was only 4%. This pattern suggests that the magnitude of the discrepancy between methods varies with the characteristics of the skill domain being measured. This principle is reinforced by the evaluation indicators used in child-friendly Health and Physical Education learning. These indicators assess students' skills, coordination, and conceptual understanding through observation during physical activities, games, and exercises (Sunardianta et al., 2024). Direct observation therefore becomes essential when the construct being assessed is actual movement performance. Consequently, a key underlying assumption that must be underscored is the presence of systematic differences between portfolio-based assessment and direct testing. This systematic difference can be explained by the fact that portfolio assessment is conducted in a familiar learning context with opportunities for multiple attempts, allowing participants to display their best performance (Fouche et al., 2021). In contrast, direct testing is conducted as a one-shot assessment intended to capture actual ability (Yu et al., 2026). In the context of physical education assessment, the assessment environment significantly influences students' performance (Otero-Saborido et al., 2021). Higher portfolio assessment scores do not necessarily indicate greater validity. Validity is not merely about measurement accuracy, but rather the extent to which score interpretations and uses can be justified for a specific purpose (Hawkins et al., 2021).

The Wilcoxon signed-rank test confirmed that differences between the two methods were statistically significant in three of the four domains: hand–eye coordination ($p < 0.001$), agility ($p < 0.001$), and endurance ($p = 0.001$). The very small p-values indicate inherent systematic differences between the two measurement methods (Warneke et al., 2024). For lower-limb muscle power, the difference was not significant ($p = 0.067$), suggesting that measurement in this domain is relatively more objective and less influenced by assessment context. This finding aligns with the principle that tests with more standardized, objective protocols tend to yield more consistent results (Nikolaidis, 2023).

Spearman correlation coefficients were positive and significant across all domains ($p < 0.01$), indicating that both methods produce relatively consistent rank-ordering. The strength of the correlations ranged from moderate ($\rho = 0.432$ for hand–eye coordination) to very strong ($\rho = 0.814$ for lower-limb muscle power). However, high correlation does not imply that the two methods are interchangeable, as correlation measures association rather than absolute agreement. Correlation analysis alone is insufficient to assess agreement between two measurement methods; more comprehensive analyses, such as ICC and Bland–Altman, are required to evaluate systematic bias (Lindberg et al., 2022; Warneke et al., 2025).

The ICC values, ranging from low to good (0.237-0.800), indicate that absolute agreement is highly domain dependent. Only lower-limb muscle power achieved "good agreement" (ICC = 0.800). Hand–eye coordination showed low ICC (0.237) with a confidence interval that included negative values, indicating a lack of meaningful agreement. In reliability assessment, ICC values

below 0.5 indicate poor reliability and are not acceptable for high-stakes decision-making (Debertin et al., 2024; Miqueleiz et al., 2025). Agility and endurance showed moderate agreement (ICC = 0.417 and 0.428), which falls within the lower bound of acceptability for university admissions selection. The low ICC values can be explained by systematic bias and high measurement variability arising from biological, instrumental, and subjective factors (Jimenez-Iglesias et al., 2024; Thron et al., 2024).

Bland–Altman analysis provides a comprehensive understanding of the magnitude, direction, and consistency of differences. Hand–eye coordination showed a significant positive bias of 5.96 points with very wide limits of agreement (−7.04 to 18.96), indicating substantial prediction error if one method is used to estimate the other. Wide limits of agreement suggest that the two methods cannot be used interchangeably at the individual level (Goosey-Tolfrey et al., 2021; Louder et al., 2021). Endurance showed a marginal bias (0.49 minutes) but extremely wide limits of agreement (−2.91 to 3.90 minutes), suggesting serious measurement inconsistency. Such high individual variability is common in endurance tests due to strong influences of motivation, pacing strategy, and environmental conditions (Ramos-Campo et al., 2026; Stjepanovic et al., 2023).

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) further confirm the substantial magnitude of prediction errors. Hand–eye coordination, with an MAE of 7.32 points (32% of the direct test mean), indicates a very high error rate. RMSE is more sensitive to outliers; therefore, large differences between MAE and RMSE may indicate the presence of extreme errors (Pang et al., 2025). Mixed-effects modeling revealed high residual variance even after accounting for differences in measurement method and inter-individual variation.

These findings have critical implications for admission policy. The low level of absolute agreement in three of the four domains indicates that portfolio scores cannot be used as a substitute for direct testing in university entrance selection. In the context of sports-related admissions, measurement methods must demonstrate high reliability and validity to ensure fair and accurate decision-making (James et al., 2024; Peng et al., 2025; Warneke et al., 2025).

This study provides comprehensive empirical evidence regarding agreement and differences between portfolio assessment and direct testing in measuring motor competence for sports science admissions. The main findings indicate systematic differences, with portfolios producing higher scores; absolute agreement ranging from low to good across domains; reasonable but insufficient convergent validity for interchangeable use; and substantial individual variability, rendering individual-level predictions unreliable.

For high-stakes decisions such as university admissions, reliance on a single method carries significant risks and limitations. A multiple-methods approach that combines the strengths of both, accompanied by continuous research on predictive validity, represents the most prudent pathway to ensuring fair, valid, and reliable admission processes in higher education in sports science.

CONCLUSION

This study demonstrates that portfolio assessment cannot yet serve as an equivalent substitute for direct motor skill testing in sports science admissions, as portfolio scores systematically overestimate performance in coordination, agility, and endurance with statistically significant differences. While correlations between methods range from moderate to very strong, the absolute agreement metrics (ICC), Bland-Altman analyses, and error measurements (MAE, MSE) reveal substantial individual score discrepancies and consistent positive bias, with good agreement observed only for lower-limb muscle power. These findings necessitate reconceptualizing portfolio assessment as a complementary rather than a standalone instrument, while maintaining standardized direct practical tests for core motor ability evaluation.

In practice, selection policies require reorientation toward hybrid assessment models, in which portfolios document participation history and achievements, while direct tests remain the primary criterion for motor competence evaluation. At the program level, post-admission motor skill mapping tests should be implemented to inform matriculation or remedial interventions, alongside strengthening internal assessment systems through clear, consistent rubrics based on

direct observation. Improving portfolio quality requires standardization through national guidelines and assessor training to enhance validity and reliability. Theoretically, this study reaffirms that motor competence, being inherently performative, demands direct observation for valid measurement and underscores the necessity of multimethod psychometric analysis when evaluating alternative assessment instruments for high-stakes selection decisions.

REFERENCES

Abudari, A. S., Al dababseh, M. F., & Al Fattah, O. A. (2022). Standard levels of physical fitness components as one of the admission indicators for students of sports excellence. *International Journal of Health Sciences*, 11641–11650. https://doi.org/10.53730/ijhs.v6nS5.12013

Carballo-Fazanes, A., Rey, E., Valentini, N. C., Rodríguez-Fernández, J. E., Varela-Casal, C., Rico-Díaz, J., Barcala-Furelos, R., & Abelairas-Gómez, C. (2021). Intra-rater (Live vs. video assessment) and inter-rater (Expert vs. novice) reliability of the test of gross motor development, third edition. *International Journal of Environmental Research and Public Health*, *18*(4), 1652. https://doi.org/10.3390/ijerph18041652

Choo, L., Novak, A., Impellizzeri, F. M., Porter, C., & Fransen, J. (2024). Skill acquisition interventions for the learning of sports-related skills: A scoping review of randomised controlled trials. *Psychology of Sport and Exercise*, *72*, 102615. https://doi.org/10.1016/j.psychsport.2024.102615

de Jong, L. H., Bok, H. G. J., Schellekens, L. H., Kremer, W. D. J., Jonker, F. H., & van der Vleuten, C. P. M. (2022). Shaping the right conditions in programmatic assessment: how quality of narrative information affects the quality of high-stakes decision-making. *BMC Medical Education*, *22*(1), 409. https://doi.org/10.1186/s12909-022-03257-2

Debertin, D., Wargel, A., & Mohr, M. (2024). Reliability of Xsens IMU-based lower extremity joint angles during in-field running. *Sensors*, *24*(3), 871. https://doi.org/10.3390/s24030871

Eddy, L. H., Bingham, D. D., Crossley, K. L., Shahid, N. F., Ellingham-Khan, M., Otteslev, A., Figueredo, N. S., Mon-Williams, M., & Hill, L. J. B. (2020). The validity and reliability of observational assessment tools available to measure fundamental movement skills in school-age children: A systematic review. *PLOS ONE*, *15*(8), e0237919. https://doi.org/10.1371/journal.pone.0237919

Fouche, I., Dison, L., Andrews, G., & Prozesky, M. (2021). Pedagogical and decolonial affordances of group portfolio assessments for learning in South African universities. *Critical Studies in Teaching and Learning*, *9*(SI). https://doi.org/10.14426/cristal.v9iSI.325

Goosey-Tolfrey, V. L., de Groot, S., Tolfrey, K., & Paulson, T. A. W. (2021). Criterion validity of a field-based assessment of aerobic capacity in wheelchair rugby athletes. *International Journal of Sports Physiology and Performance*, *16*(9), 1341–1346. https://doi.org/10.1123/ijspp.2020-0517

Hawkins, M., Elsworth, G. R., Nolte, S., & Osborne, R. H. (2021). Validity arguments for patient-reported outcomes: justifying the intended interpretation and use of data. *Journal of Patient-Reported Outcomes*, *5*(1), 64. https://doi.org/10.1186/s41687-021-00332-y

Hariadi, H., Valianto, B., Akhmad, I., & bin Syed Ali, K. S. (2025). Implementation study of authentic assessment in physical education in Indonesia and Malaysia. C*akrawala Pendidikan: Jurnal Ilmiah Pendidikan,* *44*(1), 102-115. DOI: https://doi.org/10.21831/cp.v44i1.71485

Hogan, J., Penney, D., O'Hara, E., & Scott, J. (2023). Stakeholder perceptions of the feasibility of e-portfolio-based assessment of physical literacy in primary health and physical education. *Physical Education and Sport Pedagogy*, *1*(1), 1–17. https://doi.org/10.1080/17408989.2023.2287523

Hulteen, R. M., True, L., & Kroc, E. (2023). Trust the "process"? When fundamental motor skill scores are reliably unreliable. *Measurement in Physical Education and Exercise Science*, *27*(4), 391–402. https://doi.org/10.1080/1091367X.2023.2199126

James, L. P., Haycraft, J. A. Z., Carey, D. L., & Robertson, S. J. (2024). A framework for test

measurement selection in athlete physical preparation. *Frontiers in Sports and Active Living*, *6*. https://doi.org/10.3389/fspor.2024.1406997

Jimenez-Iglesias, J., Gonzalo-Skok, O., Landi-Fernández, M., Perez-Bey, A., & Castro-Piñero, J. (2024). Age-related differences and reliability of a field-based fitness test battery in young trained footballers: The role of biological age. *Life*, *14*(11), 1448. https://doi.org/10.3390/life14111448

Kristiyandaru, A., Prakoso, B. B., Fitroni, H., Primanata, D., Hartati, S.C.Y., & Kartiko, D.C. (2023). Factors influencing assessment in higher education: Empirical evidence from physical education and fitness compulsory courses in Indonesia. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 42*(3), 617-630. DOI: https://doi.org/10.21831/cp.v42i3.60151

Lindberg, K., Solberg, P., Bjørnsen, T., Helland, C., Rønnestad, B., Thorsen Frank, M., Haugen, T., Østerås, S., Kristoffersen, M., Midttun, M., Sæland, F., Eythorsdottir, I., & Paulsen, G. (2022). Strength and power testing of athletes: Associations of common assessments over time. *International Journal of Sports Physiology and Performance*, *17*(8), 1280–1288. https://doi.org/10.1123/ijspp.2021-0557

Liu, J., & Li, Y. (2022). The visual movement analysis of physical education teaching considering the generalized hough transform model. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/3675319

Louder, T., Thompson, B. J., & Bressel, E. (2021). Association and agreement between reactive strength index and reactive strength index-modified scores. *Sports*, *9*(7), 97. https://doi.org/10.3390/sports9070097

Makaruk, H., Porter, J. M., Webster, E. K., Makaruk, B., Bodasińska, A., Zieliński, J., Tomaszewski, P., Nogal, M., Szyszka, P., Starzak, M., Śliwa, M., Banaś, M., Biegajło, M., Chaliburda, A., Gierczuk, D., Suchecki, B., Molik, B., & Sadowski, J. (2023). The fus test: a promising tool for evaluating fundamental motor skills in children and adolescents. *BMC Public Health*, *23*(1), 1912. https://doi.org/10.1186/s12889-023-16843-w

Miqueleiz, U., Aguado-Jimenez, R., Lecumberri, P., Garcia-Tabar, I., & M. Gorostiaga, E. (2025). Reliability of inertial measurement unit-based spatiotemporal and kinetic variables in endurance runners during treadmill running. *Journal of Human Kinetics*, *97*, 65–76. https://doi.org/10.5114/jhk/195895

Ng, J. L., & Button, C. (2023). Construct validation of a general movement competence assessment utilising active video gaming technology. *Frontiers in Bioengineering and Biotechnology*, *11*. https://doi.org/10.3389/fbioe.2023.1094469

Nikolaidis, P. (2023). Exercise testing and motivation. *Sci*, *5*(1), 12. https://doi.org/10.3390/sci5010012

Otero-Saborido, F. M., Torreblanca-Martínez, V., & González-Jurado, J. A. (2021). Systematic review of self-assessment in physical education. *International Journal of Environmental Research and Public Health*, *18*(2), 766. https://doi.org/10.3390/ijerph18020766

Pang, Y., Zhang, K., & Li, F. (2025). Explainable quality assessment of effective aligned skeletal representations for martial arts movements by multi-machine learning decisions. *Scientific Reports*, *15*(1), 323. https://doi.org/10.1038/s41598-024-83475-4

Peng, S., Khairani, A. Z., Rabiu Uba, A., & Yuan, F. (2025). Physical activity measurement tools among college students in intervention studies: A systematic review. *PLOS ONE*, *20*(4), e0321593. https://doi.org/10.1371/journal.pone.0321593

Penney, D., O'Hara, E., & Lund, R. (2023). Enhancing quality and equity? Performance assessment validation in examination physical education in Western Australia. *Curriculum Studies in Health and Physical Education*, *14*(3), 288–305. https://doi.org/10.1080/25742981.2022.2136007

Pool, A. O., Jaarsma, A. D. C., Driessen, E. W., & Govaerts, M. J. B. (2020). Student perspectives on competency-based portfolios: Does portfolio reflect their competence development? *Perspectives on Medical Education*, *9*(3), 166–172. https://doi.org/10.1007/S40037-020-00571-7

Ramos-Campo, D. J., Foster, C., Andreu-Caravaca, L., Rubio-Arias, J. Á., & Rosenblat, M. A. (2026). Comparative effects of pacing strategies on endurance performance: a systematic

review and network meta-analysis. *Sports Medicine*, *56*(3), 725–738. https://doi.org/10.1007/s40279-025-02367-3

Rey, E., Carballo-Fazanes, A., Varela-Casal, C., & Abelairas-Gómez, C. (2020). Reliability of the test of gross motor development: A systematic review. *PLOS ONE*, *15*(7), e0236070. https://doi.org/10.1371/journal.pone.0236070

Ross, G. B., Zhao, X., Troje, N. F., Fischer, S. L., & Graham, R. B. (2024). Assessing inter- and intra-rater reliability of movement scores and the effects of body-shape using a custom visualisation tool: an exploratory study. *BMC Sports Science, Medicine and Rehabilitation*, *16*(1), 205. https://doi.org/10.1186/s13102-024-00988-1

Stjepanovic, M., Knechtle, B., Weiss, K., Nikolaidis, P. T., Cuk, I., Thuany, M., & Sousa, C. V. (2023). Changes in pacing variation with increasing race duration in ultra-triathlon races. *Scientific Reports*, *13*(1), 3692. https://doi.org/10.1038/s41598-023-30932-1

Sunardianta, R., Prasojo, L. D., Yuliarto, H., & Firmansyah, F. (2024). Child friendly school-based learning management model for health and physical education. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 43*(2), 459-469. DOI:https://doi.org/10.21831/cp.v43i2.64652

Thron, M., Düking, P., Ruf, L., Härtel, S., Woll, A., & Altmann, S. (2024). Assessing anaerobic speed reserve: A systematic review on the validity and reliability of methods to determine maximal aerobic speed and maximal sprinting speed in running-based sports. *PLOS ONE*, *19*(1), e0296866. https://doi.org/10.1371/journal.pone.0296866

van Maarseveen, M., Leenhouts, J., de Witte, A., Flux, E., van Doorn, H., & van der Kamp, J. (2025). Enhancing affordance perception in pre-service physical education teachers: effects of content knowledge, motor experience and visual experience programs. *Frontiers in Sports and Active Living*, *7*. https://doi.org/10.3389/fspor.2025.1583448

Warneke, K., Gronwald, T., Wallot, S., Magno, A., Hillebrecht, M., & Wirth, K. (2025). Discussion on the validity of commonly used reliability indices in sports medicine and exercise science: a critical review with data simulations. *European Journal of Applied Physiology*, *125*(6), 1511–1526. https://doi.org/10.1007/s00421-025-05720-6

Warneke, K., Skratek, J., Wagner, C.-M., Wirth, K., & Keiner, M. (2024). Random measurement and prediction errors limit the practical relevance of two velocity sensors to estimate the 1RM back squat. *Frontiers in Physiology*, *15*. https://doi.org/10.3389/fphys.2024.1435103

Yang, Q., Song, M., Chen, X., Li, M., & Wang, X. (2025). The influence of linear and nonlinear pedagogy on motor skill performance: the moderating role of adaptability. *Frontiers in Psychology*, *16*(1), 1–11. https://doi.org/10.3389/fpsyg.2025.1540821

Yu, X., Huang, Z., Yang, S., Wang, Z., Ma, H., & Zhang, X. (2026). PEOAT: Personalization-guided evolutionary question assembly for one-shot adaptive testing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(33), 27978–27986. https://doi.org/10.1609/aaai.v40i33.40022