JOSSAE (Journal of Sport Science and Education)

JOSSAE

OPEN ACCESS
E-ISSN Online: 2548-4699
http://journal.unesa.ac.id/index.php/jossae/index
http://dx.doi.org/10.26740/jossae.v11n1.p12-29

KAMS: A Kids' Athletics Movement Skill Assessment Instrument-Evidence of Content Validity and Reliability in Indonesian Children Aged 10–11 Years

Ricky Wibowo[1ABCDE]*, Agus Mahendra[2AE], Yudy Hendrayana[3AE], Sucipto[4AD], Rene II Mediana Babiera[5AD]
[1,2,3,4] Universitas Pendidikan Indonesia, Indonesia
[5] University of The Immaculate Conception, Philippines

*Authors' Contribution: A – Study design; B – Data collection; C – Statistical analysis; D – Manuscript Preparation; E – Funds Collection

Correspondence: ricky_wibowo@upi.edu
Received: 20 May 2026 **Accepted**: 07 Jul 2026 **Published**: 30 Jul 2026

Abstract
The World Athletics Kids' Athletics Level 2 programme targets children aged 8–11 years through running, jumping, and throwing activities. However, no standardised and psychometrically validated instrument exists to assess Kids' Athletics movement skills among Indonesian children. This study developed and validated the Kids' Athletics Movement Skill Assessment (KAMS), comprising six sub-tests: 30 m Sprint, Hurdles Sprint, Standing Long Jump, Standing Triple Jump, Turbo Throw, and Medicine Ball Push. Following the COSMIN framework, instrument development included item generation, expert content validation, pilot testing (n = 30), and psychometric evaluation with 60 primary school students (30 boys, 30 girls). Content validity was perfect (S-CVI/Ave = 1.00). Inter-rater reliability was excellent (ICC = 0.910), intra-rater reliability ranged from 0.836–0.970, and test-retest reliability ranged from 0.750–0.871. Internal consistency was acceptable (Cronbach's α = 0.793). The mean total process score was 24.68 ± 3.33, with 83.3% of students achieving at least the Good category. Although boys showed higher throwing and jumping product scores, total process scores did not differ significantly by gender. KAMS demonstrated excellent psychometric properties and is suitable for evaluating Kids' Athletics programmes in Indonesian primary school children, with further studies recommended to establish construct and concurrent validity.

Keywords: Athletics; instrument validation; movement skills; kids; Indonesia

1. Introduction

Regular exercise enhances motor skills, allowing children to perform fundamental movements such as running, jumping, throwing and maintaining balance efficiently and confidently (Lloyd et al., 2019; Queiroz et al., 2014). This ability is essential for engaging in physical activity and maintaining lifetime physical activity habits (Valentini et al., 2017). Motor skills—which consist of locomotor, manipulative and balance skills—are positively associated with physical activity, cardiovascular health and lifelong quality of life, as children with high motor skills tend to be more physically active, and continuous physical activity in itself helps to further improve their motor skills (a positive developmental spiral) (Robinson et al., 2015; Stodden et al., 2008). In contrast, children with low motor skills may fall into a negative spiral which leads to physical inactivity and obesity (Barnett et al., 2009). In accordance with it, other recent studies confirm that low motor competence in school-age is significantly associated with low participation in sport (Barnett et al., 2024; Kokstejn et al., 2025).

Children aged 10–11 years showed a significant correlation between physical fitness, functional status and physical development, indicating dynamic and varied development of the motor system (Krasova & Krasov, 2019). A programme to develop motor skills in children aged 10–11 has shown improvements in bilateral coordination, demonstrating that targeted interventions can lead to enhanced motor skills during this period (Africa & Van Deventer, 2015). In this age group, athletic movement skills such as running, jumping and throwing are at a stage of consolidation and show a significant correlation with motor skills and physical fitness (Haga, 2008). In Indonesia, children aged 10–11 are normally in grades 4–5 of primary school — a transitional period of considerable pedagogical importance and relevance for structured interventions in athletic movement skills.

World Athletics has established a Kids' Athletics programme designed for children aged 4–7 (Level 1), 8–11 (Level 2) and 12–14 (Level 3) (World Athletics, 2022). The programme organises athletic activities across three main movement domains: running, jumping and throwing — covering eight fundamental skills: agility, balance, bilateral movements, body awareness, coordination, reaction, rhythm and spatial awareness. The programme is designed to be inclusive, adaptable to a range of facility conditions, and focuses on the development of movement skills through game-based activities (World Athletics, 2022).

Although the Kids' Athletics programme has been extensively implemented, there is a critical methodological gap: there are currently no standardised, psychometrically validated assessment tools for athletic movement skills specifically designed for the Kids' Athletics programme. World Athletics currently only provides pedagogical guidance in the form of Activity Cards, a Practitioner Handbook, and a Competition Guide, which define movement constructs — however, there is no standardized assessment rubric available which can be used for scientific research (World Athletics, 2022). The Test of Gross Motor Development-3 is a valid and reliable assessment tool for children aged 3–10 years, but it is not specifically designed for athletics programmes and does not cover Kids' Athletics events(Valentini et al., 2021). The KörperkoordinationsTest für Kinder (KTK) measures general motor coordination through four subtests of balance and coordination, but does not include sprinting, long jump or throwing, which are key components of Kids' Athletics (Vandorpe et al., 2011). Whereas the Athletic Skills Track assesses motor skills in three areas of the locomotor and stability skills in children aged 4–12 years, with good test-retest reliability (ICC: 0.80–0.88) and acceptable internal consistency (α: 0.70–0.76) (J. Hoeboer et al., 2018). However, Overall Skill Fitness Test (AST) does not cover the throwing events — one of the three main disciplines of Kids' Athletics — and is not based on the World Athletics framework. The instrument most closely aligned with it is the OSF Test within the 'Athletics for All' programme, which consists of the sprint, standing long jump and medicine ball overhead throw — three components relevant to Kids' Athletics Level 2 (Baj-Korpak et al., 2022, 2024). The OSF Test has been validated in large populations (n = 27.187) and demonstrates good reliability. However, the OSF Test does not cover the hurdles sprint and standing triple jump, and is not integrated with a process-oriented assessment of movement technique, nor is it constructed based on the World Athletics Kids' Athletics framework. Consequently, there is a need for an assessment instrument that: (1) is explicitly constructed based on the World Athletics Kids' Athletics framework; (2) covers all six subtests across the three main movement domains; (3) integrates a dual assessment of both measurable outcomes (product) and movement technique quality (process); and (4) has been validated for the Indonesian childhood population..

Another important aspect to consider in the development of this instrument is the difference in athletic performance based on gender. Evidence from previous study shows that gender differences in children's athletic performance begin to emerge at primary school age, especially after the age of 8. Papaiakovou et al. (2009) found that boys demonstrated significantly faster sprint performance than girls from the age of 8, the difference expanding as they grew older(Papaiakovou et al., 2009). Hoeboer et al. (2018) reported that boys were significantly faster than girls on all AST tracks except for the 4-

year-old age group (J. J. A. A. M. Hoeboer et al., 2018). Gallahue dan Ozmun (2012) stated that boys' superiority in throwing tasks is greater and emerges earlier than in running and jumping domains, due to a combination of morphological factors and differing motor experience (Gallahue et al., 2012). Conversely, Lopes et al. (2023) found that at the age of 10, there were no significant differences in motor competence (measured by KTK) between boys and girls engaged in structured athletics activities (Lopes et al., 2023). This pattern of varied findings indicates the need for separate normative data by gender in the KAMS instrument, as well as analyzing gender differences as part of the instrument's discriminant validity test.

The aims of this study are: (1) to develop the Kids' Athletics Movement Skill Assessment (KAMS) instrument based on the World Athletics Kids' Athletics framework (2022), referring to the OSF Test (Baj-Korpak et al., 2022, 2024); (2) to examine the content validity by seven expert panels using the Content Validity Index (Polit et al., 2007); and (3) to examine psychometric reliability (inter-rater, intra-rater, test-retest) and establish preliminary normative data (Bandung, West Java) by gender for primary school students aged 10–11 years in Indonesia. The development of KAMS is aligned with United Nations Sustainable Development Goal 3 (SDG 3: Good Health and Well-being), which emphasises the promotion of healthy lifestyles across all age groups. Physical literacy and motor skill competence in childhood are recognised as foundational to lifelong physical activity participation (Stodden et al., 2008). In the Indonesian context, evidence-based physical assessment instruments are critically needed to monitor motor development and guide programme quality.

2. Method

Design

This study employed the Instrument Development and Validation Study design, based on Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) (Mokkink et al., 2010). The COSMIN framework provides a comprehensive set of international standards for the development and evaluation of measurement properties of instruments, covering four main components: validity, reliability, responsiveness, and interpretability. The approach used is mixed-methods: qualitative for content validation by expert panel assessment, and quantitative for psychometric reliability testing and the establishment of normative data. The research was conducted in four stages designed in accordance with COSMIN recommendations: (1) instrument development, (2) content validation, (3) pilot study, and (4) psychometric testing. This research protocol has been approved by the Research Ethics Committee of Universitas Pendidikan Indonesia (Approval No. 103/UN40.K/PT.01.01/2026), and all procedures were conducted in accordance with the principles of the Declaration of Helsinki regarding research involving human subjects.

Instrument Development

The KAMS construct was developed through a systematic synthesis of four sources, prioritised according to their relevance to the Kids' Athletics construct, as summarised in Table 1. The World Athletics Kids' Athletics framework (2022) was designated as the primary source for the construct as it serves as the global normative authority in defining the domains, fundamental skills and technical specifications of the Kids' Athletics programme (World Athletics, 2022). The OSF Test within the 'Athletics for All' programme functioned as the closest validated reference — a children's athletics instrument that has been validated on a large scale (n = 27.187) and includes relevant sub-tests (sprint, standing long jump, medicine ball throw) (Baj-Korpak et al., 2022, 2024). The IAAF Kids' Athletics Practical Guide (2006) provides a historical precedent in the form of technical analyses of each event and the progression of children's skills. A process-based assessment approach using five criteria per sub-test with a 0/1 scoring system was adopted from the Fundamental Motor (Sports) Skills Test (FUS) as a methodological reference-not as a source of constructs. (Makaruk et al., 2023).

Table 1. Hierarchy of sources in the development of KAMS

No	Source	Contributions to KAMS	References
1	World Athletics KA Framework (2022)	Main source-defines 3 domains (Run/Jump/Throw), 8 basic skills, the selection of 6 subtests, and Level 2 equipment specifications	*Practitioner Handbook (2022); Activity Cards Level 2 (2022); Competition Guide (2022)*
2	OSF Test / 'Athletics for All' — (Baj-Korpak et al., 2022, 2024)	The closest validated reference-the Sprint, SLJ and Medicine Ball Throw protocols, which have been validated in n = 27.187 children participating in an athletics programme	*(Baj-Korpak et al., 2022)(Baj-Korpak et al., 2024)*
3	IAAF Kids' Athletics Practical Guide & Educational Cards (2006)	Historical precedent-technical analysis of every Kids' Athletics event and the progression of children's motor skills	*IAAF Practical Guide (2006); IAAF Educational Cards (2006)*
4	FUS Test — Makaruk et al. (2023)	Methodological references– a 5-criteria process approach for each sub-test, a 0/1 scoring system, and a video-based assessment protocol	*Makaruk et al. (2023), BMC Public Health, 23(1):1912*

The KAMS consisted of six subtests covering the three domains of Kids' Athletics Level 2: (1) 30-metre sprint (Run domain), (2) 30-metre hurdle sprint (Run domain), (3) standing long jump (Jump domain), (4) standing triple jump (Jump domain), (5) Turbo Throw (Throw domain), and (6) Medicine Ball Push (Throw domain). Each sub-test is assessed across two dimensions: (a) performance scores-directly measured outcomes (time in seconds for running sub-tests; distance in cm for jumping sub-tests; distance in metres for the throwing sub-tests) and (b) process score-the quality of movement technique assessed through five binary criteria (score 0 = not achieved; score 1 = achieved), resulting in a process score per sub-test between 0 and 5 and a total process score between 0 and 30. Equipment specifications are adapted to the age group: hurdle height 50 cm, medicine ball weight 2 kg, and the use of a junior turbo for the throwing sub-test. All these specifications refer to the World Athletics Kids' Athletics Competition Guide Level 2.

2.3 Content Validation

Content validity was carried out using the Content Validity Index (CVI) method developed by (Polit et al., 2007), the same standard used in the validation of the Fundamental Motor (Sports) Skills Test (FUS), which yielded a CVI of 1.00 across all dimensions(Makaruk et al., 2023). expert panel consisted of seven members with the following composition: two academics specialising in Athletics (holding a doctoral degree, with at least two publications in reputable journals); two internationally certified CESC World Athletics Level 2 athletics coaches; two senior physical education teachers familiar with the Kids' Athletics programme; and one expert in measurement and assessment in physical education and sport. The number of seven experts was chosen because, with this number, an I-CVI standard of ≥ 0.83 equates to agreement from at least six of the seven experts, which is a widely accepted threshold in validation research (Polit et al., 2007).

Each expert received a validation pack comprising: (1) a description of the construct being measured (Kids' Athletics Level 2 athletic movement skills); (2) the complete KAMS instrument documentation, including task descriptions, equipment specifications, and 30 process criteria (5 per subtest); (3) a rating form containing 49 items—seven items per sub-test (two general suitability items and five process criteria items) plus seven general instrument rating items; and (4) a source hierarchy reference document. Each item is assessed using a four-point scale: 1 = irrelevant, 2 = less relevant, 3 = relevant (requires minor revision), 4 = highly relevant. A notes and suggestions for improvement column is provided for each item.

The item-level content validity index (Item-CVI; I-CVI) is calculated using the formula: I-CVI = n (scores of 3 or 4) ÷ total number of experts. The overall content validity index for the instrument (Scale-CVI/Average; S-CVI/Ave) is calculated using the formula: S-CVI/Ave = ΣI-CVI ÷ total number of items. The established acceptance criteria are: I-CVI ≥ 0.83 for each item and S-CVI/Ave ≥ 0.90 for the instrument as a whole (Polit et al., 2007). Items with an I-CVI < 0.83 were revised based on the experts' notes, then discussed in a panel consensus session before being included in the final instrument. Items with an I-CVI < 0.70 were considered for deletion or complete reconstruction. Content validation was carried out in writing and independently by each expert without interaction with one another during the assessment period.

Pilot Study

The pilot study was conducted with 30 students in grades 4–5 of primary school who were not included in the main research sample, and who originated from schools other to the main research site. This selection was intended to avoid a practice effect on the main sample. Two trained raters conducted the pilot study together and systematically recorded issues identified across eight aspects: (1) time taken per subtest and total session duration; (2) clarity of verbal instructions for 9–10-year-old children; (3) suitability of equipment specifications; (4) clarity and understanding of the five process criteria by the assessors; (5) the quality of video recordings from the designated camera positions; (6) the safety of administering each sub-test; (7) the students' responses and enthusiasm; and (8) unexpected logistical obstacles. The results of the pilot study were used as the basis for the final revision of the instrument prior to the psychometric testing. Students who participated in the pilot study were not included in the main sample.

Pilot study findings resulted in three revisions to the KAMS instrument prior to main data collection: (1) the verbal instruction script for the Hurdles Sprint sub-test was simplified after children showed difficulty understanding the original phrasing; (2) the camera position for the Standing Triple Jump was adjusted from a 45-degree angle to a direct lateral view to ensure all three landing phases were fully visible; and (3) the inter-subtest rest interval was extended from 45 to 60 seconds based on observed fatigue indicators in younger participants. All other aspects — including equipment specifications, safety protocols, process criteria clarity, and student engagement — were confirmed adequate without revision. Mean session duration was 62 ± 8 minutes for all six sub-tests.

Participants

Table 2 presents the study participants. The participants are 10-11-year-old 4th and 5th grade elementary school students from public schools in Bandung, West Java, Indonesia. The 10–11-year-old age group was treated as a single homogeneous group based on empirical evidence showing that performance differences between these two age groups are not statistically significant on comparable motor skills instruments (J. J. A. A. M. Hoeboer et al., 2018). The sampling technique employed purposive sampling based on age, gender, and school accessibility. Inclusion criteria included: (1) active students enrolled at the corresponding school; (2) aged 10–11 years at the time of data collection; (3) have no active musculoskeletal injuries that could affect motor performance; (4) have obtained informed consent from a parent or guardian; and (5) declared their agreement to participate. Exclusion criteria included acute health conditions on the day of the test.

Table 2. Participants

Parameter	Specification	Justification
Total main sample	n = 60 students in grades 4–5 of primary school (30 boys + 30 girls)	Sufficient for ICC with a 95% confidence interval and gender-specific tests; following AST, n=30 per group (Hoeboer et al., 2017); Makaruk et al. (2023)
Age group	Ages 10–11 (4th–5th grade): treated as a single homogeneous group	Hoeboer et al. (2018): the difference of 9–10 years on the AST-3 was not significant ($p > 0.05$ for both boys and girls); WA Level 2 covers 8–11 years within the same level
Gender	Equal numbers: 30 boys and 30 girls	Allows for an independent t-test comparing boys and girls as a measure of discriminant validity; consistent with the AST and OSF Test, which report separate norms for each gender
Location	2 primary schools in one town; purposive sampling	Representative of the local target population
Inclusion criteria	Active students; healthy; without musculoskeletal injuries	Participant safety
Reliability sub-sample	n = 30 from the main group (±15 boys; ±15 girls) — measured twice at intervals of 2 weeks	For the ICC test-retest; ensure an even distribution in order for the ICC to be representative of all sub-groups

The total main sample for the psychometric testing comprised 60 students, consisting of 30 boys and 30 girls. A balanced gender composition was required to enable the analysis of gender differences as part of the discriminant validity testing. A sample size of n = 60 is considered adequate based on existing precedents : Hoeboer et al. (2018) used n = 30 per age group to calculate the ICC reliability of the AST with a reasonable 95% CI, whilst Makaruk et al. (2023) used n = 81 per age group for the reliability test of the FUS Test (J. Hoeboer et al., 2018; Makaruk et al., 2023). For the test-retest reliability sub-sample, 30 students were randomly selected from the main sample (±15 boys, ±15 girls), who were re-measured after an interval of 2 weeks. All research procedures, including the obtaining of informed consent and child assent, were carried out in accordance with the research ethics standards established by the institution.

Data Collection Procedure

Prior to the official data collection, the two raters involved in the study underwent a structured training session lasting at least eight hours. The training session comprised: (1) independent study of the KAMS instrument manual and all process criteria for each subtest; (2) independently viewing and assessing demonstration videos for each subtest; (3) a calibration discussion session to agree on the interpretation of potentially ambiguous criteria; and (4) a reliability test using 10–15 videos of different subjects, with a target ICC of ≥ 0.80, prior to the commencement of formal testing. This assessor training approach follows the protocol used in the FUS Test (Makaruk et al., 2023), where intensive training was shown to yield inter-rater ICCs between 0.97 and 0.98.

The data collection sessions were conducted on the school sports field during physical education lessons. The sequence of sub-tests was standardised for all participants: Sprint → Hurdles → Standing Long Jump → Standing Triple Jump → Turbo Throw → Medicine Ball Throw. Before each sub-test, the assessor provided standard verbal instructions and a live demonstration in accordance with a pre-determined script. Each participant was given one familiarisation trial to understand the task and safety procedures, followed by two official trials (T1 and T2) with a minimum one-minute rest period between trials. All attempts are recorded using a video camera from a pre-determined position for each sub-test — the camera position is determined to ensure the student's entire body is clearly visible and all process criteria can be observed. Assessment of process criteria (C1–C5) is carried out from the video recording using slow-motion playback for sub-tests requiring detailed observation of movement technique. Both assessors evaluate independently without being able to view each other's assessment sheets.

For the intra-rater reliability test, Rater A re-scored the same video recordings after a minimum interval of four weeks from the first assessment, without access to the results of their initial assessment. This four-week interval was chosen to minimise the effect of recalling previous scores, following the procedure used by Makaruk et al. (2023), which resulted in an intra-rater ICC between 0.95 and 0.99. For the test-retest reliability test, a sub-sample of n = 30 students was re-measured by Rater A after an interval of two weeks (second measurement session), under identical testing conditions: the same time, place, equipment, rater, and protocol. An interval of two weeks was chosen so that the students' physical condition would not change significantly, yet was long enough to avoid the practice effect from the first session.

Statistical Analysis

Content validity measured using the I-CVI for each item and the S-CVI/Ave overall score for the instrument, as described in Section 2.3. The acceptance thresholds used were I-CVI ≥ 0.83 and S-CVI/Ave ≥ 0.90 (Polit et al., 2007), consistent with the standards used in the FUS Test (CVI = 1.00 for all aspects) (Makaruk et al., 2023).

Inter-Rater Reliability. Inter-rater reliability among scores from Rater A and Rater B was quantified with Intraclass Correlation Coefficient (ICC) model ICC(2,1)—two-way mixed, single measures, absolute agreement—which is the recommended model for reliability studies where raters are specifically selected and generalization is limited to the same raters (Koo & Li, 2016). In addition to the ICC with a 95% confidence interval (CI), the following were also calculated: Cohen's kappa ($\varkappa$) coefficient per subtest as a measure of categorical reliability, interpreted using the Landis & Koch classification (1977: $\varkappa < 0.40$ = weak; 0.40–0.60 = moderate; 0.61–0.80 = substantial; > 0.80 = almost perfect); as well as the percentage of agreement, calculated as the proportion of subjects for whom both assessors gave identical total process scores per sub-test. Interpretation of ICCs follows the guidelines of Koo & Mae (2016): ICC < 0.50 = poor; 0.50–0.74 = moderate; 0.75–0.89 = good; ≥ 0.90 = very good.

Intra-Rater Reliability. The internal consistency of Rater A's assessments was tested using the same ICC(2.1) between the Session 1 scores (first assessment) and the re-scoring scores (second assessment after four weeks). No systematic bias was detected using a one-sample t-test on the difference between the two assessments (mean difference = 0 as the null hypothesis), and proportional bias was examined using Pearson's correlation between the means and the difference between the two assessments (Makaruk et al., 2023).

Test-Retest Reliability. The stability of KAMS scores over time was assessed using ICC(2,1) between Session 1 and Session 2 scores for a sub-sample of n = 30. The following were calculated: (a) Standard Error of Measurement ($SEM = SD \times \sqrt{(1 - ICC)}$); (b) $MDC_{95} = SEM \times 1.96 \times \sqrt{2}$, representing the smallest change interpretable as genuine rather than measurement noise; and (c) Bland-Altman analysis for each subtest, presenting the limits of agreement (LOA = mean difference ± 1.96 × SD of difference). Bland-Altman plots for Hurdles Sprint (lowest ICC = 0.750) and MBall (ICC = 0.769) are presented in Figure 1 and Figure 2 respectively; plots for the remaining four subtests are provided as Supplementary Figures S1–S4.

Internal Consistency. Internal consistency between the five process criteria (C1–C5) within each subtest, as well as between subtests within each domain (Run: T1+T2; Jump: T3+T4; Throw: T5+T6), was assessed using Cronbach's alpha. A value of $\alpha \geq 0.70$ was set as the acceptable threshold, following the AST, which achieved $\alpha = 0.70$–0.76 (J. J. A. A. M. Hoeboer et al., 2018). It should be noted that in observation-based assessment instruments with process criteria that measure different aspects of a single complex movement, a moderate Cronbach's alpha value (0.60–0.80) is methodologically acceptable because each criterion is designed to measure different aspects of movement technique, rather than a fully homogeneous construct (Makaruk et al., 2023).

Descriptive Statistics and Normative Data. Descriptive statistics were calculated for all main variables: mean (M), standard deviation (SD), minimum value, maximum value, and the P25, P50 (median) and P75 percentiles. Descriptive statistics are presented for the entire sample ($n = 60$) and separately by gender ($n = 30$ males; $n = 30$ females). The normality of the data distribution was tested using the Shapiro-Wilk test, given that the sample size per gender was < 50 (Field et al., 2018).

Gender Differences (Discriminant Validity). Prior to group comparison, each variable's normality was checked with the Shapiro-Wilk test (recommended for $n < 50$ per group; Field, 2018). An independent samples t-test was used for normally distributed variables ($p > 0.05$); while if normality was not met ($p \leq 0.05$), the Mann-Whitney U test was performed as a non-parametric alternative. The effect size was calculated by Cohen's d (t-test) or the biserial rank correlation r (Mann-Whitney U), defined as: trivial < 0.20; small 0.20–0.49; moderate 0.50–0.79; large 0.80–1.29; very large ≥ 1.30 (Cohen, 1988). Bonferroni correction was applied across six comparisons (adjusted $\alpha = 0.008$).

3. Result

Participant Characteristics

The participants in this study were 60 students (30 boys; 30 girls) from a public elementary school in Bandung, West Java, Indonesia, aged 10–11 years. All participants met the inclusion criteria. Two trained raters completed a minimum of eight hours of calibration prior to data collection. A sub-sample of 30 students (±15 boys; ±15 girls) participated in a retest session with an interval of 2 weeks.

Normality Testing

Prior to gender comparison analyses, normality was assessed using the Shapiro-Wilk test for each variable separately in boys and girls ($n = 30$ each). Product scores for SLJ, STJ, Turbo Throw, and MBall were normally distributed in both groups (all $p > .05$) and were compared using independent samples t-tests. 30 m Sprint product (Boys: $W = 0.798$, $p < .001$) and Hurdles Sprint product (Boys: $W = 0.917$, $p = .023$) violated normality in the boys group; Mann-Whitney U tests were applied for these variables. All six process scores violated normality in at least one group (all $p \leq .05$) and were analysed using Mann-Whitney U.

Content Validity

Table 3 presents the results of the assessment by seven expert panels. All 49 items achieved an I-CVI of 1.00 — all seven experts rated each item as 3 or 4, with no scores below 3. The overall S-CVI/Ave = 1.00, exceeding the minimum threshold of ≥ 0.90 (Polit et al., 2007) and identical to the FUS Test (Makaruk et al., 2023). The average score per expert ranged from 3.22 to 4.00 (grand mean = 3.78 ± 0.18). All 49 items were accepted without revision and the KAMS instrument was validated in terms of content.

Table 3. Content validity index (CVI) results of the KAMS instrument (N experts = 7, N items = 49)

Item Group	N Item	Mean ±SD	I-CVI	S-CVI	Status
Sub-testt 1: 30 m Sprint	7	3.79 ±0.15	1.00	1.00	✓ Valid
Sub-testt 2: Hurdle Sprint	7	3.80 ±0.18	1.00	1.00	✓ Valid
Sub-testt 3: SLJ	7	3.80 ±0.15	1.00	1.00	✓ Valid
Sub-testt 4: STJ	7	3.80 ±0.07	1.00	1.00	✓ Valid
Sub-testt 5: Turbo Throw	7	3.84 ±0.09	1.00	1.00	✓ Valid
Sub-testt 6: MBall	7	3.73 ±0.11	1.00	1.00	✓ Valid
Overall Assessment U1–U7	7	3.73 ±0.10	1.00	1.00	✓ Valid
Total (49 item)	49	3.78 ±0.18	1.00	1.00	✓ Valid

Note: I-CVI = Item-Content Validity Index; S-CVI = Scale-Content Validity Index/Average. Expert grand mean: A1=3.22; A2=3.80; A3=4.00; A4=3.67; A5=3.96; A6=3.84; A7=3.98. Criteria: I-CVI ≥ 0.83; S-CVI/Ave ≥ 0.90 (Polit et al., 2007).

Inter-rater Reliability

Table 4 presents the inter-rater ICCs (N = 60). ICC values ranged from 0.875 to 0.932, with an overall ICC of 0.910 [95% CI: 0.854, 0.945] (excellent). Sub-test values: STJ = 0.932 [0.889, 0.959] (excellent); Turbo Throw = 0.909 [0.852, 0.945] (excellent); 30 m Sprint = 0.906 [0.847, 0.943] (excellent); MBall = 0.899 [0.836, 0.938] (good); Hurdles Sprint = 0.877 [0.802, 0.925] (good); SLJ = 0.875 [0.799, 0.923] (good). Percentage agreement ranged from 83.3% to 88.3%; SEM = 0.241–0.288 points.

Table 4. Inter-rater reliability of the KAMS (N = 60)

Sub-test	n	Rater A M(SD)	Rater B M(SD)	ICC	SEM	% Agree	Category
T1: 30 m Sprint	60	4.20 (0.80)	4.15 (0.78)	0.906 [0.847, 0.943]	0.241	88.3%	Excellent
T2: Hurdles Sprint	60	3.85 (0.82)	3.78 (0.83)	0.877 [0.802, 0.925]	0.288	83.3%	Good
T3: SLJ	60	3.65 (0.82)	3.68 (0.81)	0.875 [0.799, 0.923]	0.288	83.3%	Good
T4: STJ	60	3.53 (0.91)	3.62 (0.94)	0.932 [0.889, 0.959]	0.240	88.3%	Excellent
T5: Turbo Throw	60	3.93 (0.88)	4.03 (0.82)	0.909 [0.852, 0.945]	0.257	86.7%	Excellent
T6: MBall	60	4.38 (0.83)	4.35 (0.80)	0.899 [0.836, 0.938]	0.257	86.7%	Good
Total	60	—	—	0.910 [0.854, 0.945]	—	86.1%	Excellent

Note: ICC = ICC(2.1) two-way mixed, absolute agreement; SEM = Standard Error of Measurement; % Agree = percentage of identical scores. Categories: Good = 0.75–0.89; Excellent ≥ 0.90 (Koo & Mae, 2016).

Intra-rater Reliability

Table 5 presents the intra-rater ICCs (N = 60). ICCs ranged from 0.836 to 0.970. STJ = 0.970 [95% CI: 0.950, 0.982] (excellent); Turbo Throw = 0.930 [0.886, 0.958] (excellent); Hurdles Sprint = 0.855 [0.769, 0.911] (good); 30 m Sprint = 0.848 [0.758, 0.906] (good); MBall = 0.848 [0.758, 0.906] (good); SLJ = 0.836 [0.740, 0.899] (good). Mean differences between sessions were all < 0.11 points, indicating no systematic bias (Bland & Altman, 1986).

Table 5. Intra-rater reliability for Rater A: Session 1 vs. re-scoring (N = 60, 4-week interval)

Sub-test	n	Session 1 M(SD)	Re-scoring M(SD)	ICC	% Agree	Mean Diff	Category
T1: 30 m Sprint	60	4.20 (0.80)	4.15 (0.76)	0.848 [0.758, 0.906]	81.7%	+0.05	Good
T2: Hurdles Sprint	60	3.85 (0.82)	3.75 (0.97)	0.855 [0.769, 0.911]	76.7%	+0.10	Good
T3: SLJ	60	3.65 (0.82)	3.72 (0.74)	0.836 [0.740, 0.899]	80.0%	−0.07	Good
T4: STJ	60	3.53 (0.91)	3.52 (0.91)	0.970 [0.950, 0.982]	95.0%	+0.02	**Excellent**
T5: Turbo Throw	60	3.93 (0.88)	3.92 (0.94)	0.930 [0.886, 0.958]	88.3%	+0.02	**Excellent**
T6: MBall	60	4.38 (0.83)	4.30 (0.72)	0.848 [0.758, 0.906]	81.7%	+0.08	Good

Note: Mean Diff = mean for Session 1 minus the re-scoring (positive = Session 1 is higher). Categories by Koo & Mae (2016).

Test-Retest Reliability

Table 6 presents the test-retest ICCs (n = 30 sub-samples, 1–2 week interval). All six subtests achieved the good category (ICC = 0.750–0.871). Turbo Throw = 0.871 [95% CI: 0.747, 0.936]; 30 m Sprint = 0.861 [0.729, 0.931]; SLJ = 0.834 [0.680, 0.917]; STJ = 0.802 [0.625, 0.901]; MBall = 0.769 [0.569, 0.883]; Hurdles Sprint = 0.750 [0.538, 0.873]. MDC_{95} ranged from 0.795 to 1.285 points: a change of ≥ 1 point per sub-test between sessions can be interpreted as a genuine change beyond measurement error (Hoeboer et al., 2017).

Table 6. Test-retest reliability of the KAMS (n = 30 sub-samples, 1–2 week interval)

Sub-test	n	Session 1 M(SD)	Session 2 M(SD)	ICC	SEM	MDC_{95}	±LOA	Category
T1: 30 m Sprint	30	4.30 (0.79)	4.23 (0.90)	0.861 [0.729, 0.931]	0.314	0.870	±0.882	Good
T2: Hurdles Sprint	30	3.90 (0.96)	3.93 (0.91)	0.750 [0.538, 0.873]	0.463	1.285	±1.311	Good

Sub-test	n	Session 1 M(SD)	Session 2 M(SD)	ICC	SEM	MDC_{95}	±LOA	Category
T3: SLJ	30	3.50 (0.90)	3.53 (0.78)	0.834 [0.680, 0.917]	0.340	0.941	±0.961	Good
T4: STJ	30	3.37 (0.85)	3.40 (0.89)	0.802 [0.625, 0.901]	0.385	1.068	±1.090	Good
T5: Turbo	30	3.83 (0.87)	3.87 (0.73)	0.871 [0.747, 0.936]	0.287	0.795	±0.811	Good
T6: MBall	30	4.30 (0.79)	4.23 (0.73)	0.769 [0.569, 0.883]	0.363	1.007	±1.021	Good

Note. SEM = SD × √(1 − ICC); MDC_{95} = SEM × 1.96 × √2 (the smallest change that can be interpreted as clinically meaningful at a 95% confidence level); ±LOA = Limits of Agreement from the Bland-Altman plot (mean difference ± 1.96 × SD of differences). All subtests met the Good standard (ICC ≥ 0.75; Koo & Mae, 2016).

Internal Consistency

Cronbach's alpha for the six subtest total scores, treated as scale items, resulted in α = 0.793 (acceptable to good) (Jum C. Nunnally & Ira H. Bernstein, 1994)). The alpha per domain ranged from 0.337 to 0.436, and the alpha for the 30 binary items was 0.604. The alpha per subtest (K1–K5 within a single subtest) was low to negative — a phenomenon discussed in the Discussion (Table 7).

Table 6. Internal consistency (cronbach's alpha) of the KAMS (N = 60)

Unit of Analysis	N Item	Cronbach's α	N	Interpretation
Per Sub-tes (K1–K5 in each sub-test)				
T1: 30 m sprint	5	−0.029	60	*Design artefact — see Discussion*
T2: Hurdles Sprint	5	−0.044	60	*Design artefact — see Discussion*
T3: SLJ	5	−0.167	60	*Design artefact — see Discussion*
T4: STJ	5	−0.074	60	*Design artefact — see Discussion*
T5: Turbo	5	0.148	60	*Low — see Discussion*
T6: MBall	5	0.052	60	*Low — see Discussion*
Per Domain (K1–K5 all sub-test in the domain)				
Run (T1+T2, 10 item)	10	0.384	60	*Below the threshold — expected by design*
Jump (T3+T4, 10 item)	10	0.337	60	*Below the threshold — expected by design*
Throw (T5+T6, 10 item)	10	0.436	60	*Below the threshold — expected by design*
All 30 binary items	30	0.604	60	*Moderate*
6 total subtest scores (scale levels)	**6**	**0.793**	**60**	**✓ Acceptable-to-Good (α ≥ 0.70)**

Note. The scale alpha (6 subtests in total, α=0.793) is a psychometric value relevant to the interpretation of the KAMS as a single scale. Low/negative alpha values for K1–K5 are artefacts of the process-criterion-based instrument design (Carmines & Zeller, 1979; Streiner, 2003). Standard: α ≥ 0.70 = acceptable (Jum C. Nunnally & Ira H. Bernstein, 1994).

Descriptive Statistics and Initial Normative Data

Table 8 shows the descriptive statistics for the KAMS product and process scores. The average 30 sprint time was 6.14 ± 0.86 seconds; the average hurdles sprint time was 8.03 ± 1.32 seconds. The average standing long jump was 126.55 ± 21.27 cm; Standing Triple Jump 317.14 ± 66.70 cm (Min = 154.1 cm; Max = 459.0 cm; P50 = 325 cm) — approximately 2.5× the SLJ value, consistent with theoretical estimates for this age group and confirming that the three-phase hop-step-jump protocol was performed correctly. The average Turbo distance was 7.23 ± 2.70 m and the Medicine Ball Push distance was 2.63 ± 0.63 m. The average total KAMS process score was 24.68 ± 3.33 (range 15–30; P50 = 25; Table 6).

Category distribution: 30.0% Very Good (≥ 27; n = 18), 53.3% Good (22–26; n = 32), 15.0% Fair (16–21; n = 9), and 1.7% Poor (10–15; n = 1). No students were in the Very Poor category. A total of 83.3% of students achieved at least the Good category.

Table 7. Descriptive statistics for KAMS product and process scores (N = 60)

Variable	N	M	SD	Min	Max	P25	P50	P75	Unit
Run Domain									
30 m sprint — Product	60	6.14	0.86	5.07	9.52	5.61	6.00	6.48	*seconds*
30 m sprint — Process	60	4.22	0.76	2	5	4	4	5	*score*
Hurdles — Process — Product	60	8.03	1.32	5.73	11.61	7.18	7.84	8.77	*seconds*
Hurdles — Process — Process	60	3.88	0.83	1	5	4	4	4	*score*
Jump Domain									
SLJ — Product	60	126.55	21.27	67.0	170.0	110.0	130.0	140.0	*cm*
SLJ — Process	60	3.97	0.82	1	5	4	4	4	*score*
STJ — Product	60	317.14	66.70	154.1	459.0	275.0	325.0	364.0	*cm*
STJ — Process	60	3.90	0.86	1	5	3	4	4	*score*
Throw Domain									
Turbo — Product	60	7.23	2.70	2.80	13.60	5.00	7.25	9.10	*m*
Turbo — Process	60	4.18	0.81	2	5	4	4	5	*score*
MBall — Product	60	2.63	0.63	1.30	4.00	2.10	2.70	3.10	*m*
MBall — Process	60	4.53	0.65	3	5	4	5	5	*score*
Total KAMS Process	**60**	**24.68**	**3.33**	**15**	**30**	**23**	**25**	**27**	*score*

Note. Data are derived from the domain assessment sheets (Run, Jump, Throw). M = Mean; SD = Standard Deviation; P25/P50/P75 = 25th/50th/75th percentile. Process scores on a scale of 0–5 per subtest; total 0–30. SLJ = Standing Long Jump.

Differences by Gender (Discriminant Validity)

Shapiro-Wilk tests indicated normality violations in at least one gender group for all six process scores and for 30 m Sprint and Hurdles Sprint product scores ($p \leq .05$); Mann-Whitney U tests were therefore applied for these variables. Product scores for SLJ (Boys: W = 0.973, p = .634; Girls: W = 0.963, p = .368), STJ (Boys: W = 0.974, p = .643; Girls: W = 0.982, p = .865), Turbo (Boys: W = 0.980, p = .817; Girls: W = 0.961, p = .331), and MBall (Boys: W = 0.972, p = .589; Girls: W = 0.984, p = .913) satisfied normality in both groups and were compared using independent samples t-tests. Effect sizes are reported as Cohen's d (t-test variables) or rank-biserial correlation r (Mann-Whitney U variables).

Table 9 presents a comparison of KAMS scores between boys and girls. Normality testing (Section 3.2 above) informed test selection. For normally distributed product variables, independent samples t-tests were used; Mann-Whitney U was used for all process scores and non-normally distributed product variables. Significant large-to-very-large differences were found in product scores: Turbo (Boys = 9.07 ± 2.19 m vs Girls = 5.39 ± 1.74 m; $t(58) = 7.22$, $p < .001$, d = 1.86); MBall ($t(58) = 4.78$, $p < .001$, d = 1.23); STJ ($t(58) = 2.91$, p = .005, d = 0.75); SLJ ($t(58) = 2.49$, p = .015, d = 0.64). Hurdles product showed boys faster (Mann-Whitney U = 289.5, p = .004, d = −0.77).

Process scores showed no significant differences by gender across all subtests (Mann-Whitney U tests; all effect sizes trivial-to-small, $|d| \leq 0.40$, Bonferroni-adjusted $\alpha = 0.008$). Total KAMS process score: Boys = 24.87 ± 2.39 vs Girls = 24.50 ± 4.09 (U = 431.5, p = .609, rank-biserial r = 0.04 — trivial), confirming KAMS is gender-neutral in assessing movement technique quality.

Table 8. Comparison of KAMS scores between boys (n=30) and girls (n=30)

Variable	Boys M(SD)	Girls M(SD)	t / U	d	Interpretation
Run Domain					
Sprint Product (dtk)	6.13 (1.08)	6.16 (0.60)	U = 439.0	−0.04	Trivial/ns[a]
Sprint Process	4.37 (0.81)	4.07 (0.69)	U = 368.0	0.40	Small/ns
Hurdles Product (dtk)	7.56 (1.22)	8.51 (1.26)	U = 289.5**	−0.77	Medium★
Hurdles Process	3.93 (0.74)	3.83 (0.91)	0.47	0.12	Trivial/ns
Jump Domain					
SLJ Product (cm)	133.10 (19.13)	120.00 (21.59)	2.49*	0.64	Medium★ (Boys > Girls)
SLJ Process	3.93 (0.52)	4.00 (1.05)	−0.31	−0.08	Trivial/ns
STJ Product (cm)	340.76 (63.23)	293.52 (62.44)	2.91**	0.75	Medium★ (Boys > Girls)
STJ Process	3.83 (0.65)	3.97 (1.03)	−0.60	−0.16	Trivial/ns
Throw Domain					
Turbo Product (m)	9.07 (2.19)	5.39 (1.74)	7.22***	1.86	Very large ★ (Boys > Girls)
Turbo Process	4.23 (0.63)	4.13 (0.97)	0.47	0.12	Trivial/ns
MBall Product (m)	2.96 (0.57)	2.30 (0.51)	4.78***	1.23	Large ★ (Boys > Girls)
MBall Process	4.57 (0.63)	4.50 (0.68)	0.39	0.10	Trivial/ns
Total Process	**24.87 (2.39)**	**24.50 (4.09)**	**0.42**	**0.11**	**Trivial/ns — no bias gender ✓**

Note. d = Cohen's d for t-test variables; rank-biserial r for Mann-Whitney U variables (Sprint, Hurdles, all Process scores). Effect size interpretation: trivial < 0.20; small 0.20–0.49; medium 0.50–0.79; large 0.80–1.29; very large ≥ 1.30 (Cohen, 1988). Bonferroni correction applied: α = 0.008. ns = not significant; * p < .05; ** p < .01; *** p < .001. ★ = significant after Bonferroni correction. [a] = Mann-Whitney U test applied (normality violated); t / U column: t values for parametric tests, U values for non-parametric tests. Sprint Product (s) = seconds.

4. Discussion

Content Validity: Full Consensus on the Hierarchical Foundation of Sources

An S-CVI/Ave score of 1.00 across all 49 items represents the highest level of content validity within the CVI methodology (Polit et al., 2007), identical to the FUS Test (Makaruk et al., 2023). The unanimous consensus of the seven expert panels reflects the strength of the source hierarchy: the World Athletics Kids' Athletics framework (2022), as the primary construct source, possesses global normative authority that is directly recognised by the entire panel. As stated by Streiner and Norman, high content validity is achieved when the items of an instrument are based on sources whose authority is recognised by experts in the relevant domain (Streiner et al., 2024).

The inclusion of the OSF Test (n = 27.187 children) as a validated reference provided convincing empirical evidence for the expert panel to accept the operationalisation of the construct (Baj-Korpak et al., 2022). The grand mean of the expert ratings was 3.78 (no score < 3 on any item) and the lowest score by Expert A1 (mean = 3.22) reflects academic rigour that actually reinforces the objectivity of the validation process. Yusoff (2019) emphasises that transparency regarding the source of the construct is a prerequisite for obtaining consistent assessments from a panel of experts — a condition that is optimally fulfilled in this study.

Reliability Profile: Satisfactory Psychometric Properties

The reliability profile of the KAMS indicates satisfactory psychometric validity across all measurement dimensions. The inter-rater ICC = 0.910 (excellent) exceeds the minimum standard of ICC ≥ 0.80 (Altman, 1991; Koo & Mae, 2016) and approaches the FUS Test's performance (ICC = 0.97–0.98) (Makaruk et al., 2023). An agreement percentage of 83.3–88.3% with a SEM of 0.241–0.288 points confirms that inter-rater differences are practically minimal. The intra-rater ICC ranged from 0.836 to 0.970 with a mean difference of < 0.11 points across all subtests, demonstrating the consistency of Rater A's assessments over time without systematic bias. These findings are consistent with similar sports movement assessment instruments reporting intra-rater ICCs of 0.80–0.95 (Lund & Tannehill, 2015).

Regarding test-retest reliability, all six sub-tests met the 'good' category (ICC = 0.750–0.871), exceeding the COSMIN minimum standard of ICC ≥ 0.70 (Mokkink et al., 2010) and comparable to the AST test-retest ICC of 0.80–0.88 (J. Hoeboer et al., 2018). MDC_{95} values ranging from 0.795 to 1.285 points provide concrete practical guidance: a change in process score of ≥ 1 point per sub-test between two measurements can be interpreted as a significant change exceeding the margin of measurement error — supporting the use of KAMS in the pre-post evaluation of the Kids' Athletics programme.

Low Alpha by Subtest: An Expected Design Artifact

Low or negative Cronbach's alpha values for K1–K5 within a single subtest are an expected design artefact, not a psychometric failure. Carmines and Zeller (1979) and Streiner (2003) explain that low alpha occurs when items are designed to measure different aspects of a single complex skill — a situation that is indeed intentional in the KAMS. Items K1–K5 in each subtest measure technically distinct components: for example, in the Sprint subtest, K1 = starting position, K2 = acceleration, K3 = posture, K4 = arm coordination, K5 = stride pattern (G.Carmines & A.Zeller, 1979; Streiner et al., 2024) A student may master one aspect but not the other, so the low inter-K correlation is by design. An identical phenomenon has been reported in TGMD-3 (Ulrich, D. A. 2019) dan FUS Test (Makaruk et al., 2023).

The psychometrically meaningful alpha is at the scale level: the six subtest total scores yielded an α of 0.793 (acceptable to good), exceeding the AST range of 0.70–0.76 (J. J. A. A. M. Hoeboer et al., 2018). This confirms that the total KAMS score (0–30) is a cohesive measure of athletic motor skills and can be interpreted as a single, meaningful construct.

Normative Profile and International Comparison

The initial normative data provide an initial overview of the athletic movement skills profile of Kids' Athletics participants among Indonesian primary school pupils. The average standing long jump (SLJ) of 126.55 cm falls below the European P50 norm of approximately 140–155 cm for 10–11-year-olds (Thomas et al., 2020) but is close to the Latin American population P50 norm ≈ 109–120 cm (Ramírez-vélez et al., 2017), reinforcing the urgency for representative local norms. The average Standing Triple Jump = 317.14 cm (≈ 2.5× SLJ) aligns with theoretical estimates and confirms that the three-phase hop-step-jump protocol was correctly executed. The mean turbo throw = 7.23 m and medicine ball push = 2.63 m fall within a reasonable range for this age group compared to OSF Test data (Baj-Korpak et al., 2022).

The distribution of 83.3% of students in the Good or Very Good categories is consistent with previous findings that participants in structured athletics programmes outperform the general population in all components of athletic movement skills. One student (1.7%) in the Poor category reflects reasonable individual variation (Baj-Korpak et al., 2022).

Concurrent validity of KAMS against established motor skill instruments was not assessed in the present study, as this constitutes a separate validation phase. However, the theoretical basis for convergent validity is well-founded. The KAMS jumping domain — specifically the Standing Long Jump — shares direct procedural overlap with the TGMD-3 horizontal jump subtest (Ulrich, 2019), while the running domain parallels sprint and agility components of the Athletic Skills Track (Hoeboer et al., 2018). The throwing domain corresponds conceptually to the medicine ball push of the OSF Test (Baj-Korpak et al., 2022) and the two-hand overarm throw of the TGMD-3. On this basis, moderate-to-strong correlations ($r = 0.50$–0.75) between KAMS domain scores and corresponding criterion instrument subscales would be theoretically expected. Establishing concurrent validity — via simultaneous administration of KAMS and TGMD-3 or AST on a new sample of ≥ 60 children aged 10–11 years — is identified as the highest-priority next step in the full KAMS validation programme.

Product Divergence — Process Convergence: Implications for Discriminant Validity

The most significant pattern was a large divergence in product scores in the throwing domain ($d = 1.23$–1.86 — large to very large) and a moderate divergence in the jumping domain ($d = 0.64$–0.75) between genders, accompanied by consistent convergence in process scores (all $|d| \leq 0.40$). Males' superiority in throwing product scores is consistent with (Gallahue et al., 2012) who attributed this advantage to a combination of morphological factors and motor experience, and with (J. J. A. A. M. Hoeboer et al., 2018) who reported that males were faster across all AST tracks.

The non-significant difference in process scores — including the total KAMS process score ($d = 0.11$ — trivial) — demonstrates that the instrument measures movement technique quality in a gender-neutral manner. This finding is consistent with (Lopes et al., 2023) who found no difference in motor competence (KTK) between 10-year-old boys and girls participating in structured athletics training. A critical practical implication: KAMS process scores are a gender-fair measure of technical evaluation and can be used without distinguishing between gender-specific norms, whilst product scores must always be interpreted using separate norms for each gender. These findings also underscore the novelty of the KAMS: the first instrument within the context of Kids' Athletics Indonesia to explicitly separate the assessment of physical capacity (product) from the quality of movement technique (process), providing an evaluation dimension not available in either the OSF Test or the AST.

Limitations of the Study

Firstly, the limited sample from schools in Bandung restricts generalisability of normative data to Indonesia's diverse national population. Secondly, concurrent validity against the TGMD-3 (Ulrich, 2019), KTK (Vandorpe et al., 2011), or AST (Hoeboer et al., 2018) has not yet been established — this is the most critical gap in the current validation evidence and should be addressed as a priority in future research. Thirdly, maturity status was not controlled, although its influence at ages 10–11 years is relatively small (MIRWALD et al., 2002).

5. Conclusion and Recommendation

The KAMS has been successfully developed and validated as an assessment tool for athletic movement skills in Kids' Athletics , based on the World Athletics Kids' Athletics framework (2022), for primary school students aged 10–11 in Indonesia. This instrument has been shown to possess perfect content validity based on expert panel assessment, comprehensive and satisfactory reliability profiles across all measurement dimensions (inter-rater, intra-rater, and test-retest), as well as acceptable internal consistency at the scale level. The KAMS instrument has also been shown to be non-gender-biased in measuring the quality of movement techniques, rendering it a fair and objective evaluation tool. Initial normative data indicate that the majority of primary school students in Bandung City have achieved a good level of mastery of athletic movement skills. The KAMS is recommended as an evidence-based evaluation tool for the Kids' Athletics programme in Indonesian primary schools, with priority for further research including concurrent validity testing, the development of multi-provincial national norms, and validation across other age groups within the Kids' Athletics spectrum.

Acknowledgement

The authors would like to express their sincere gratitude to all parties who contributed to this research. This study was supported by the Indonesian Education Scholarship (BPI Program), funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia in collaboration with the Indonesia Endowment Fund for Education (LPDP).

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article. The authors also confirmed that this paper is free of plagiarism.

References

Africa, E. K., & Van Deventer, K. J. (2015). Balance and bilateral skills of selected previously disadvantaged children aged 9 to 12 years. *South African Journal for Research in Sport, Physical Education and Recreation*. https://www.scopus.com/pages/publications/84949549956?origin=scopusAI

Altman, D. G. (1990). Practical Statistics for Medical Research. *Practical Statistics for Medical Research*. https://doi.org/10.1201/9780429258589

Baj-Korpak, J., Stelmach, M. J., Zaworski, K., Lichograj, P., & Wochna, M. (2022). Assessment of Motor Abilities and Physical Fitness in Youth in the Context of Talent Identification—OSF Test. *International Journal of Environmental Research and Public Health*, *19*(21), 14303. https://doi.org/10.3390/IJERPH192114303/S1

Baj-Korpak, J., Zaworski, K., Wochna, M., Chmara, S., & Stelmach, M. J. (2024). Physical fitness and body build parameters of children and adolescents participating in the physical activity promotion programme "Athletics for all!" *Archives of Public Health 2024 82:1*, *82*(1), 206-. https://doi.org/10.1186/S13690-024-01436-9

Barnett, L. M., van Beurden, E., Morgan, P. J., Brooks, L. O., & Beard, J. R. (2009). Childhood Motor Skill Proficiency as a Predictor of Adolescent Physical Activity. *Journal of Adolescent Health*, *44*(3), 252–259. https://doi.org/10.1016/j.jadohealth.2008.07.004

Barnett, L. M., Verswijveren, S. J. J. M., Colvin, B., Lubans, D. R., Telford, R. M., Lander, N. J., Schott, N., Tietjens, M., Hesketh, K. D., Morgan, P. J., Hinkley, T., Downing, K. L., Telford, R. D., Cohen, K. E., Ridgers, N. D., & Abbott, G. (2024). Motor skill competence and moderate- and vigorous-intensity physical activity: a linear and non-linear cross-sectional analysis of eight pooled trials. *International Journal of Behavioral Nutrition and Physical Activity*, *21*(1). https://doi.org/10.1186/s12966-023-01546-7

Field, A. P., Field, & P., A. (2018). Discovering statistics using IBM SPSS statistics. *Discovering Statistics Using IBM SPSS Statistics*, 510–531. https://docs.edtechhub.org/lib/CFJAF8HE

Gallahue, D. L., Ozmun, J. C., & Goodway, Jackie. (2012). *Understanding motor development : infants, children, adolescents, adults*. 461. https://books.google.com/books/about/Understanding_Motor_Development.html?hl=id&id=DVBhpwAACAAJ

G.Carmines, E., & A.Zeller, R. (1979). Reliability and Validity Assessment. *Reliability and Validity Assessment*. https://doi.org/10.4135/9781412985642

Haga, M. (2008). The relationship between physical fitness and motor competence in children. *Child: Care, Health and Development*, *34*(3), 329–334. https://doi.org/10.1111/j.1365-2214.2008.00814.x

Hoeboer, J. J. A. A. M., Ongena, G., Krijger-Hombergen, M., Stolk, E., Savelsbergh, G. J. P., & de Vries, S. I. (2018). The Athletic Skills Track: Age- and gender-related normative values of a motor skills test for 4- to 12-year-old children. *Journal of Science and Medicine in Sport*, *21*(9), 975–979. https://doi.org/10.1016/j.jsams.2018.01.014

Hoeboer, J., Krijger-Hombergen, M., Savelsbergh, G., & De Vries, S. (2018). Reliability and concurrent validity of a motor skill competence test among 4- to 12-year old children. *Journal of Sports Sciences*, *36*(14), 1607–1613. https://doi.org/10.1080/02640414.2017.1406296

Jum C. Nunnally, & Ira H. Bernstein. (1994). Bernstein: psychometric theory. *McGraw-Hill, New York*, (1994), 2015–2018. https://search.worldcat.org/title/28221417

Kokstejn, J., Grobar, M., Vampola, J., & Musalek, M. (2025). Why Motor Competence Matters: Fundamental Movement Skills and Their Role in Promoting Physical Activity and Health in Czech Children Aged 9–10 Years. *Journal of Functional Morphology and Kinesiology*, *10*(3). https://doi.org/10.3390/jfmk10030258

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Krasova, I., & Krasov, O. (2019). Informative significance of indicators of physical preparedness and morphofunctional state in the structure of the motor system of 9–10 years old pupils. *Slobozhanskyi Herald of Science and Sport*, *2019*(1), 18–24. https://doi.org/10.15391/snsv.2019-1.003

Lloyd, R. S., Moeskops, S., & Granacher, U. (2019). Motor skill training for young athletes. In *Strength and Conditioning for Young Athletes: Science and Application* (pp. 103–130). Taylor and Francis. https://doi.org/10.4324/9781351115346-6

Lopes, N., Matos, R., Amaro, N., Coelho, L., Antunes, R., Jacinto, M., Rodrigues, F., Monteiro, D., & Ibáñez, S. (2023). Motor competence of 10 years old children with different athletics practice years. *Retos*, *50*, 599–604. https://doi.org/10.47197/RETOS.V50.99333

Lund, J. Lea., & Tannehill, Deborah. (2015). *Standards-based physical education curriculum development*. 422.

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), 1–18. https://doi.org/10.1186/s12889-023-16843-w

MIRWALD, R. L., G. BAXTER-JONES, A. D., BAILEY, D. A., & BEUNEN, G. P. (2002). An assessment of maturity from anthropometric measurements. *Medicine and Science in Sports and Exercise*, *34*(4), 689–694. https://doi.org/10.1097/00005768-200204000-00020

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & De Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *19*(4), 539–549. https://doi.org/10.1007/S11136-010-9606-8

Papaiakovou, G., Giannakos, A., Michailidis, C., Patikas, D., Bassa, E., Kalopisis, V., Anthrakidis, N., & Kotzamanidis, C. (2009). The effect of chronological age and gender on the development of sprint performance during childhood and puberty. *Journal of Strength and Conditioning Research*, *23*(9), 2568–2573. https://doi.org/10.1519/JSC.0B013E3181C0D8EC

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467. https://doi.org/10.1002/NUR.20199

Queiroz, D. D. R., Ré, A. H. N., Henrique, R. D. S., Moura, M. D. S., & Cattuzzo, M. T. (2014). Participation in sports practice and motor competence in preschoolers. *Motriz. Revista de Educacao Fisica*, *20*(1), 26–32. https://doi.org/10.1590/S1980-65742014000100004

Ramírez-vélez, R., Rosario, U., Correa-bautista, J. E., Rosario, U., & García-hermoso, A. (2017). *Self-Rated Health Status and Cardiorespiratory Fitness in a Sample of Schoolchildren from Self-Rated Health Status and Cardiorespiratory Fitness in a Sample of Schoolchildren from Bogot á ,*. (August). https://doi.org/10.3390/ijerph14090952

Robinson, L. E., Stodden, D. F., Barnett, L. M., Lopes, V. P., Logan, S. W., Rodrigues, L. P., & D'Hondt, E. (2015). Motor Competence and its Effect on Positive Developmental Trajectories of Health. *Sports Medicine (Auckland, N.Z.)*, *45*(9), 1273–1284. https://doi.org/10.1007/S40279-015-0351-6

Stodden, D. F., Langendorfer, S. J., Goodway, J. D., Roberton, M. A., Rudisill, M. E., Garcia, C., & Garcia, L. E. (2008). A Developmental Perspective on the Role of Motor Skill Competence in Physical Activity: An Emergent Relationship. *Quest*, *60*(2), 290–306. https://doi.org/10.1080/00336297.2008.10483582

Streiner, D. L. ., Norman, G. R. ., & Cairney, John. (2024). *Health measurement scales : a practical guide to their development and use*. https://books.google.com/books/about/Health_Measurement_Scales.html?hl=id&id=pb30EAAAQBAJ

Thomas, E., Petrigna, L., Tabacchi, G., Teixeira, E., Pajaujiene, S., Sturm, D. J., Nese Sahin, F., Gómez-López, M., Pausic, J., Paoli, A., Alesi, M., & Bianco, A. (2020). Percentile values of the standing broad jump in children and adolescents aged 6-18 years old. *European Journal of Translational Myology*, *30*(2), 9050. https://doi.org/10.4081/EJTM.2019.9050

Valentini, N. C., Nobre, G. C., Zanella, L. W., Pereira, K. G., Albuquerque, M. R., Rudisill, M. E., Valentini, N. C., Nobre, G. C., Zanella, L. W., Pereira, K. G., Albuquerque, M. R., & Rudisill, M. E. (2021). Test of Gross Motor Development–3 Validity and Reliability: A Screening Form. *Journal of Motor Learning and Development*, *9*(3), 438–455. https://doi.org/10.1123/JMLD.2020-0061

Valentini, N. C., Pierosan, L., Rudisill, M. E., & Hastie, P. A. (2017). Mastery and exercise play interventions: motor skill development and verbal recall of children with and without disabilities. *Physical Education and Sport Pedagogy*, *22*(4), 349–363. https://doi.org/10.1080/17408989.2016.1241223

Vandorpe, B., Vandendriessche, J., Lefevre, J., Pion, J., Vaeyens, R., Matthys, S., Philippaerts, R., & Lenoir, M. (2011). The KörperkoordinationsTest für Kinder: Reference values and suitability for 6-12-year-old children in Flanders. *Scandinavian Journal of Medicine and Science in Sports*, *21*(3), 378–388. https://doi.org/10.1111/J.1600-0838.2009.01067.X;WGROUP:STRING:PUBLICATION

World Athletics. (2022). *Kids Athletics | World Athletics*. https://worldathletics.org/kids-athletics