Science and Medicine in Football

ISSN: 2473-3938 (Print) 2473-4446 (Online) Journal homepage: www.tandfonline.com/journals/rsmf20

Evaluation of a goalkeeper-specific motor coordination assessment in youth football

Fynn Bergmann, Florian Schultz, Job Fransen & Oliver Höner

To cite this article: Fynn Bergmann, Florian Schultz, Job Fransen & Oliver Höner (28 Nov 2024): Evaluation of a goalkeeper-specific motor coordination assessment in youth football, Science and Medicine in Football, DOI: 10.1080/24733938.2024.2429486

To link to this article: https://doi.org/10.1080/24733938.2024.2429486

SCIENCE AND MEDICINE IN FOOTBALL
https://doi.org/10.1080/24733938.2024.2429486

Routledge
Taylor & Francis Group

Evaluation of a goalkeeper-specific motor coordination assessment in youth football

Fynn Bergmann [a], Florian Schultz [a], Job Fransen [b] and Oliver Höner [a]

[a]Institute of Sports Science, Eberhard Karls University Tübingen, Tübingen, Germany; [b]School of Allied Health, Exercise and Sports Sciences, Charles Sturt University, Port Macquarie, Australia

ABSTRACT

Talent identification and development (TID) in football can be enhanced through objective assessments of talent predictors. Yet, available instruments rarely consider the unique demands of goalkeepers (GKs). During early phases of talent development, considering a GK's giftedness relating to, for example, different abilities (e.g. motor coordination), can complement views on highly specialized GK-specific attributes (e.g. technical skills). Therefore, this study aimed to evaluate a GK-specific motor coordination assessment to support TID in football at early developmental phases. Six tests were designed to assess ball control relevant to GKs. Their content and face validity were confirmed by independent experts ($N = 8$). The assessment was evaluated with GKs selected for the German Football Association's TID program (U12-U15; $N = 120$). This study examined the assessment's test–retest reliability and agreement, the structural validity, and the concurrent validity. The overall test score demonstrated good test–retest reliability, although some individual tests showed lower coefficients. Additionally, limited agreement between repeated measurements due to considerable measurement error as well as issues with the structural validity of the test battery were identified. Nevertheless, the findings support the assessment's concurrent validity as higher-rated (i.e. more talented) GKs outperformed lower-rated individuals. Yet, the test battery's diagnostic accuracy is not high enough to justify individual decisions for talent selection. Overall, these findings support the consideration of motor coordination as a talent predictor in youth GKs and emphasize the assessment's potential to enhance coaches' evaluations. The identified psychometric weaknesses in some individual tests provide impetus to further optimize the test battery.

ARTICLE HISTORY
Accepted 11 November 2024

KEYWORDS
Talent assessment; talent predictors; test development; playing position; Soccer

The goalkeeper (GK) plays a crucial role for a football team's success and should therefore receive concerted attention along the talent pathway (Vahia and Kelly 2024). Recent literature reviews on GK-specific game demands emphasize the need for multidimensional ability- and skill-related attributes such as quickness, power, as well as several GK-specific technical and tactical skills (West 2018; Otte et al. 2023). This empirical knowledge about the demands imposed upon GKs in the game contributes to a sophisticated understanding of the determinants required for high-level performance. However, this body of research hardly informs about the attributes, which should be considered as *predictors of goalkeeping talent* at a young age. Given this limited knowledge about talent predictors in GKs, current talent identification and development (TID) practices remain largely a product of the expertise and experience of practitioners (Vahia and Kelly 2024). This paucity of research evidence alongside the fact that well-established practical approaches to identifying and developing talented GKs remain under-researched, severely limits our understanding of how these processes can be optimized.

This limited understanding opposes trends regarding the growing empirical insights on talent in sports in general (Baker et al. 2020; for a review) and in football more specifically (Williams et al. 2020; for a review). In football, these studies have informed sport practice pertaining to, for example, the consideration of several prognostically relevant talent predictors in outfield players. In this regard, research demonstrates higher predictive power of football-specific compared to general motor attributes in youth players already selected (Sieghartsleitner, Zuber, Zibung, Charbonnet et al. 2019). Furthermore, different studies report the best predictions if subjective ratings by practitioners were combined with objective assessments (e.g., motor tests; Sieghartsleitner, Zuber, Zibung, Conzelmann 2019; Dugdale et al. 2020; Höner et al. 2021). Yet, the available research also emphasizes the need to consider talent predictors in the light of different developmental phases given a variety of mediating and moderating factors (e.g., training history; access to high-quality coaching; Williams et al. 2020).

The relevance of accounting for developmental phases is also emphasized by the *Differentiated Model of Giftedness and Talent* (*DMGT*; Gagné 2021). According to the *DMGT*, a person's (innate) *giftedness* refers to the possession of exceptional abilities (e.g., coordination) which are relevant for a certain field (e.g., goalkeeping). These abilities form the foundation for the development of *talent*, characterized by a person's systematically developed, extraordinary competencies (see also Preckel et al. 2020[1]). Consequently, in early developmental phases, talent assessments may not only focus on highly specialized, sport-specific attributes,

CONTACT Fynn Bergmann fynn.bergmann@uni-tuebingen.de Institute of Sports Science, Eberhard Karls University Tübingen, Wilhelmstrasse 124, Tübingen 72074, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/24733938.2024.2429486

whose development is largely related to the training history (Vandorpe et al. 2012). The additional consideration of abilities as indicators of a GK's giftedness and, thus, the potential to develop high levels of GK-specific competencies can add valuable information.

Talent research in the field of football goalkeeping

The relatively scarce talent research focusing on GKs is, for example, due to the far lower number of GKs compared to outfield players in football squads. For this reason, GKs were seemingly often deliberately excluded or omitted in previous talent research (Vahia and Kelly 2024). One exception presents the study by Gil et al. (2014), who examined talent predictors in GKs and observed that 9- to 10-year-old GKs were more likely to be selected when being taller and performing better in generic, physical tests (e.g., jump tests). Another study by Höner et al. (in revision) demonstrated that position-independent motor tests assessing technical skills (e.g., passing and ball control) and speed abilities (e.g., linear-sprint) had predictive power for selection into a German youth academy (YA) 3 years after the assessment for a sample of pre-selected GKs aged U12 to U15. In this study, however, additionally conducted subjective coach ratings of GK-specific technical and tactical skills showed even higher predictive power in univariate analyses. Yet, multivariate analyses indicated that these coach ratings discriminated less between distinct GK-specific technical and tactical skill domains. Thus, the subjective ratings represented accurate, but seemingly general judgments of a GK's talent. These findings hint at the contribution of objective motor tests for more differentiated assessments of specific talent predictors (see also Höner et al. 2021 for a discussion).

These studies further concluded that TID could be optimized if objective assessments would better account for GK-specific demands. Such objective tools become even more relevant if no GK-experts were available for subjective ratings, and for generating reference values that allow for a systematic comparison of GKs' talent in large-scale TID programs. There are, however, only a few scientifically evaluated *GK-specific assessments* available in the literature. Performance in these assessments was often related to both GK-specific technical skills (e.g., diving) and physical attributes (e.g., jump performance; Knoop et al. 2013; Rebelo-Gonçalves et al. 2016). Although scientific evaluations with (late) adolescent GKs underline the assessments' concurrent validity in these age groups, their predictive validity in early developmental phases is unknown and the test performance seems highly associated with GKs' anthropometrics and maturation. Another challenge with these assessments in practice is their comparably complex test setups, which limit their practical feasibility.

Consequently, there is a lack of objective and practically feasible assessments of talent predictors suitable to support GKs' TID in early developmental phases. At these early time points, also considering abilities that underpin GK-specific developmental processes seem relevant (Preckel et al. 2020; Gagné 2021). These attributes could indicate a youth GK's potential to develop high levels of GK-specific competence when being systematically promoted within TID programs (e.g., position-specific practice; coaching by GK experts, etc.).

Motor coordination as a potential talent predictor in youth sports

The applied and academic literature on football GKs highlights that motor coordination is one relevant ability, and further identifies domain-specific attributes related to coordination (e.g., 'ball handling'; 'eye-hand coordination'; Busch 2017; Elleray 2021; Otte et al. 2023). That motor coordination can serve as a predictor of talent is further supported by empirical studies across various sports, which describe it as one indicator of an individual's potential for sport-specific skill development (Vandorpe et al. 2012; Pion, Fransen, et al. 2015, Pion, Lenior, et al., 2015; Rommers et al. 2019). Therefore, assessments of motor coordination may reveal the potential to support GKs' talent identification in practice.

In the scientific literature, manifold theories and taxonomies on motor abilities have been proposed (e.g., general motor ability vs. specificity hypotheses; Haibach et al. 2018; Hands et al. 2018). At present, multidimensional conceptualizations of motor abilities, one of which is motor coordination, seem most accepted (in detail, e.g., Lämmle et al. 2010). Motor coordination relates to the patterning of head, body, or limb motions relative to the patterning of environmental objects and events (Turvey 1990; Magill and Anderson 2021). Utesch and Bardid (2019) describe motor coordination as an underlying mechanism of motor competence and, thus, fundamental for goal-directed movements.

This variety of phenomenological and empirical concepts also led to several factorial structures of motor coordination assessments (e.g., Lämmle et al. 2010). These assessments can broadly be categorized by the types of motor tasks they use, which are based on either static environments (e.g., jumping sideways over a line; Kiphard and Schilling 2007) or dynamic environments (e.g., controlling a ball; Faber et al. 2015). Particularly in dynamic environments, participants need to continuously adjust their movements based on environmental information (e.g., ball movements) to optimally solve a motor task. This categorization shows parallels to assessments of higher-order constructs like general motor competence, often represented by latent factors termed *locomotion* and *object control* (Aadland et al. 2022; gross motor function and ball control, Faber et al. 2015; self-movement and object-movement; Herrmann and Seelig 2017; locomotor and ball skills; Garn and Webster 2021).

Despite the use of such test batteries within a deficit-oriented approach (e.g., detecting motor coordination issues; Fransen et al. 2014; Coppens et al. 2021), there is some evidence from studies conducted in different sports that motor coordination assessments can support the detection and identification of sporting talent. A review by O'Brien-Smith et al. (2019) indicates that the 'Körper Koordinationstest für Kinder' (KTK; Kiphard and Schilling 2007), which is based on motor tasks in static environments, can explain part of the variance between youth athletes competing at different performance levels across

sporting domains. Furthermore, a *cross-sectional study* indicated the concurrent validity of ball-related tests (i.e., motor tasks in dynamic environments) in table tennis (Faber et al. 2014).

In *prospective study designs*, it has been found that test items of the KTK relating to maintaining balance possess predictive power for child gymnasts' participation in an elite talent pathway 2 years after the assessment (Vandorpe et al. 2012). Similarly, Deprez et al. (2014) confirmed the two-year predictive validity of the KTK test items 'jumping sideways', 'moving sideways', and 'balancing backward' in early to mid-adolescent Belgian soccer players. Two further studies partially confirmed the predictive validity of ball-related 'eye-hand coordination' and 'ball skills' tests in table tennis players below the age of 11 (Faber et al. 2016, 2023). A *longitudinal study* further demonstrated that coordination tests with and without ball were among the best predictors in identifying handball talent (Matthys et al. 2013).

In summary, previous research indicates that assessments of motor coordination can aid the detection and identification of sport-specific talent, especially when applied during the early developmental phases. However, most studies similarly emphasize the importance of considering specific aspects of motor coordination, taking into account the unique demands of each sport.

Development of a GK-specific motor coordination assessment

Accounting for football GKs' highly specialized positional demands, Bergmann et al. (2021) presented a GK-specific motor coordination assessment that was developed based on a co-design involving practitioners and researchers. Specifically, an expert panel, consisting of GK experts working for the German Football Association's (DFB) TID program (i.e., full-time practitioners with several years of experience) and full-time researchers providing scientific support for the program, developed the assessment through a five-step process (in detail *Supplement I*): Initially, the experts inductively developed a test battery (step 1). A first pilot study (step 2) demonstrated good to very good procedural objectivity ($.90 \leq r_{xy} \leq .99$) and moderate to good test–retest reliability ($.66 \leq r_{xy} \leq .89$). In a subsequent pilot study evaluating the concurrent validity (step 3), it was found that the individual tests demonstrated diagnostic power in distinguishing between pre-selected GKs who were rated as having different levels of talent by experts ($.10 \leq \eta_p^2 \leq .22$). Despite these promising findings found within the two pilot studies (i.e., steps 2 & 3; Bergmann et al., 2021), practical experience when conducting the assessment and statistical results encouraged the expert panel to further optimize the test battery (e.g., practical feasibility, test demands; step 4). Finally, these further developed tests underwent an independent expert rating that demonstrated the test battery's content and face validity, as well as the practical feasibility (step 5; in detail *Supplement 1*). The six test items finally included in the test battery are presented in Table 1 (Bergmann et al. 2024).

The present study

The present study aimed to evaluate the motor coordination assessment designed to support TID in youth football GKs. The evaluation was conducted at both a global level (i.e., the overall test performance) and with respect to each test item. These processes followed the consensus-based standards for the selection of health status measurement instruments (COSMIN), particularly guidelines for evaluating instrument's measurement properties (Mokkink et al. 2010a, 2020). Guided by the COSMIN taxonomy on instrument's measurement properties (Mokkink et al. 2010b), this study pursues three objectives addressing the assessment's reliability (objective 1) and validity (objectives 2 & 3):

(1) Assessing the test battery's *internal consistency*, *test-retest-reliability* (i.e., 'relative reliability'), and *agreement* across repeated measurements (i.e., 'absolute reliability') to estimate the degree of measurement error in the test data.
(2) Exploring the *structural validity* of the test battery to identify potentially underlying dimensions as the theoretical construct of motor coordination is inconsistently defined, assessed, and discussed in the literature.
(3) Evaluating the *concurrent validity* in terms of the assessment's accuracy in identifying goalkeeping talent which is essential to support TID in practice.

Methods

Sample

To evaluate the psychometric properties of the test battery, male U12 to U15 GKs were tested ($N = 120$; in detail Table 2). All GKs were selected for a base camp (BC; 'competence center') of the DFB TID program located in two different regional football associations, namely *Fußball- und Leichtathletik Verband Westfalen* ('Westphalia'; $n = 69$; 25 different BCs) and *Württembergischer Fußballverband* ('Württemberg'; $n = 51$; 17 different BCs). At these BCs, approximately the best 4% of GKs in their respective age were offered one additional practice session per week complementing their regular clubs' activities (Kelly et al. 2024). Both regions are split up within the DFB's TID program into two 'coordination areas', each supervised by a full-time coach coordinator. The decision to assess GKs from these two regions was justified, as they were represented by four comparably large coordination areas (between 11 and 16 BCs), thus contributing to achieving large sample sizes given the commonly low number of GKs at each BC. Additionally, the two coordination areas in each region prepare their players for one regional association squad, further enhancing the comparability of these two regions (see also 'criterion variable'). As the data were collected at two different time points in Westphalia (June 2022; Season 2021/2022) and Württemberg (October 2022; Season 2022/2023), two different birth cohorts were assessed (Westphalia: 2007–2010; Württemberg: 2008–2011). However, no significant difference in the subsamples' age was found ($t(118) = 1.697$, $p = .53$) so that these two time points had no consequences for the statistical analyses.

Table 1. Overview of the six individual tests included to the test battery.

	1. Juggling-Ball-Change (JBC)	2. Double-Ball-Bouncing (DBB)
	  	  
Equipment	• Adhesive tape (alternative: hall line) • Two juggling balls in different shades/colours (70 g, 62 mm) • Smartphone/tablet as stopwatch (with three-second countdown)	• 2 age-appropriate footballs • 8 cones (alternative: basketball center circle) • Adhesive tape • Stopwatch
Set-up	The goalkeeper stands behind an indoor boundary line (alternative: tape line) and holds one juggling ball in each hand, which are different in colour.	The goalkeeper stands in the centre of a circle (diameter: 3.6 metres). The starting point at the center of the circle is marked by a 1-meter-long strip of adhesive tape.
Procedure	The goalkeeper throws the juggling balls from a shoulder-width arm position simultaneously from one hand to the other. A pre-determined ball (colour-coded) is always thrown over the other ball.	The goalkeeper initiates the test independently by simultaneously bouncing both balls on the indoor floor. His task is to bounce the balls above his own waist level using the same hand for each ball, without securing them or leaving the circle.
Familiarization	• 4 throws	• 4× simultaneous bouncing
Duration	• 2 attempts each lasting for 30 seconds	• 2 attempts, which end as soon as at least one ball can no longer be controlled, are secured or after max. 45 seconds
Rating	• 1 point is awarded if both balls are caught • All points scored during the 30-second test period are summed	• 1 point if both bounced balls are touched again above the goalkeeper's waist with the respective bouncing hand • All points scored during the test period are summed
Errors Patterns & Consequences	• If at least one ball is dropped, no point is awarded for the throw. The test continues. • If the goalkeeper maintains an arm position narrower than shoulder-width, indicated by the verbal cue from the test supervisor ('shoulder!') no points are awarded. The test continues.	• At least one ball can no longer be controlled (attempt ends). • At least one ball is secured (e.g., pressed against the body). • The goalkeeper or a ball touches the ground outside the circle. • At least one ball bounced below the hip (no point, test continues).
	3. Ball-Wall-Rebounder – one Ball (BWR I)	**4. Ball-Wall-Rebounder – two Balls (BWR II)**
	  	  
Equipment	• Adhesive tape • Ruler or measuring tape • At least 4 age-appropriate footballs • 1 vaulting box or similar equipment for spare balls • Smartphone/tablet as stopwatch (with three-second countdown)	• Adhesive tape • Ruler or measuring tape • At least 6 age-appropriate footballs • 1 vaulting box or similar equipment for spare balls • Smartphone/tablet as stopwatch (with three-second countdown)
Set-up	The marking line is adhered to the floor, parallel to the indoor wall, with a distance of 1.4 metres.	The marking line is adhered to the floor, parallel to the indoor wall, with a distance of 1.2 metres.
Procedure	The goalkeeper throws a football using both hands from below the elbows against the indoor wall and catches it with both hands above his own elbows. This sequence is repeated as many times as possible within the testing time.	The goalkeeper simultaneously throws two footballs against the indoor wall and catches them with the same hand. This sequence is repeated as many times as possible within the testing time.
Familiarization	• 4 throws	• 4 throws
Duration	• 2 attempts each lasting for 30 seconds	• 2 attempts each lasting for 30 seconds
Rating	• 1 point is awarded if the ball is thrown below the elbows and caught above the elbows • All points scored during the 20-second test period are summed	• 1 point is awarded if both balls are caught • All points scored during the 30-second test period are summed

(*Continued*)

Table 1. (Continued).

Errors Patterns & Consequences	• The ball is dropped (no point for the throw, the test continues) • The goalkeeper crosses the marking line when throwing or catching (no point for the throw, the test continues)	• At least one ball is dropped (no point for the throw, the test continues) • The goalkeeper crosses the marking line when throwing or catching (no point for the throw, the test continues)
	5. Throwing-Sitting-Catching (TSC)	**6. Bouncing-Turning-Catching (BTC)**
	  	  
Equipment	• Adhesive tape (alternatively indoor line) • 1 age-appropriate footballs • At least 15 × 15 metres of free indoor space	• Adhesive tape • 1 age-appropriate football • Smartphone/tablet as a stopwatch (with three-second countdown)
Set-up	A free hall area of at least 15 × 15metres is required. On one side of this area, an indoor line (alternative: line of adhesive tape) is marked as the starting position.	A strip of adhesive tape line (length: 1 meter) is adhered to the indoor floor. At each end of the line, a strip of adhesive tape measuring 50 cm is marked at a right angle to the line.
Procedure	The goalkeeper throws a football into the air using both hands, quickly sits down on the indoor floor, stands back up, and catches the ball as high in the air as possible. After each throw, the goalkeeper has time to recover before starting the next throw. The quality of each attempt is assessed using a four-point scale.	The goalkeeper bounces the ball in front of him on the indoor floor, rotates 360 degrees around his own axis and attempts to catch the ball in the starting position (i.e., standing with his feet to the right and left of the marking line). Then, he performs the same sequence with a 360-degree rotation to the other side and repeats this as many times as possible.
Familiarization	• 2 throws	• four trials (2× rotation to each side)
Duration	• 1 round á 5 throws (the points of the 5 throws are summed)	• 2 attempts each lasting for 15 seconds
Rating	The ball is … … not caught or caught while still sitting down (0 points) … caught below the chin while standing (1 point) … caught above the chin while standing (2 points) … caught above the head in a jump (3 points)	• 1 point is awarded if the goalkeeper catches the ball after the turn back in the starting position. • All points scored in the 15-second test period are summed
Errors Patterns & Consequences	• Goalkeeper did not touch the hall floor when sitting (no point) • Ball touches indoor ceiling (repetition of throw)	• Goalkeeper is not in starting position when catching the ball • Goalkeeper turns to the wrong side (no point, test continues) • Goalkeeper loses the ball or falls down (no point, test continues)

Table 2. Sample characteristics separated by the four investigated age groups and two different regions.

	Total Sample				Westphalia				Württemberg			
	N	Age (yrs.)	Height (cm)	Weight (kg)	*n*	Age (yrs.)	Height (cm)	Weight (kg)	*n*	Age (yrs.)	Height (cm)	Weight (kg)
Overall	120	12.93 ± 1.20	162.93 ± 11.67	50.32 ± 12.06	69	13.12 ± 1.13	164.72 ± 10.77	52.38 ± 11.61	51	12.65 ± 1.21	160.49 ± 12.47	47.53 ± 12.22
U12	42	11.70 ± 0.41	154.04 ± 8.12	42.16 ± 7.36	25	11.95 ± 0.32	156.20 ± 7.46	44.55 ± 6.94	17	11.34 ± 0.21	150.87 ± 8.02	38.66 ± 6.69
U13	34	12.72 ± 0.41	160.63 ± 8.79	47.51 ± 8.29	19	12.97 ± 0.27	162.46 ± 6.14	49.62 ± 5.96	15	12.39 ± 0.31	158.31 ± 11.11	44.85 ± 10.32
U14	21	13.84 ± 0.43	171.36 ± 8.81	56.70 ± 8.72	15	14.05 ± 0.29	174.34 ± 6.66	59.69 ± 7.55	6	13.29 ± 0.11	163.90 ± 9.66	49.21 ± 7.09
U15	23	14.61 ± 0.42	174.84 ± 7.12	63.45 ± 12.38	10	14.95 ± 0.32	175.91 ± 8.77	66.25 ± 15.70	13	14.36 ± 0.29	174.02 ± 5.79	61.46 ± 9.24

Before entering the DFB TID program, written informed consent for the recording and scientific use of each individual's data was provided by a legal guardian/next of kin. The first author's university's ethics department approved the use of the data for this study.

Data collection

The data collection was embedded into GK-specific training camps of the DFB, including sufficient recovery breaks. Guided by the detailed test manual (Bergmann et al. 2024; for an overview of all individual tests see *Table I*), research assistants, trained and supervised by a member of the research team, conducted all data collection. Before participating in the assessment, all participants filled out a one-page questionnaire covering questions on personal information (e.g., name, date of birth), training history (e.g., GK experience), and their current club. Additionally, anthropometric data were assessed. Weight was measured with calibrated scales (Seca 813 electronic flat scale) to the nearest 0.1 kg, and height was measured to the nearest 0.1 cm with a fixed stadiometer (Seca 213 portable stadiometer).

All GKs completed the whole test battery twice (i.e., test–retest design) with a recovery break in between. The tests were conducted in a fixed circle during both the test (T1) and retest (T2) to ensure a comparable time between both measurement points. With an average completion time of approx. 45 min. for the entire assessment per group, ensuring sufficient and equally comparable recovery, the test–retest interval for the same test was approx. 60 minutes.

Across all tests six tests included in the test battery, the number of points collected represents the metric test outcome. Similar to previous studies using ball-related tests (Faber et al. 2015), the best out of two attempts was considered for analyses except for the TSC in which the sum of points out of five trials represents the test outcome. A familiarization phase before each test should ensure participants' comprehension, but least affect their baseline performance (i.e., a person's initial performance level on a given task; Baltes 1987).

Criterion variable

All tested GKs were assigned into three groups by expert GK-coaches who were not involved in the motor testing. They assigned GKs to *group A* who possess sufficient talent to be considered for selection into a regional association squad, representing the next higher selection level in the DFB TID program. *Group B* includes GKs who were appraised to be among the top 4% of GKs in their respective age cohort and were hence considered to possess sufficient talent to remain at their BC. *Group C* includes GKs who were considered to possess little talent and hence their deselection from the TID program should be contemplated. The coach rating was conducted after GKs participated in training camps that serve as measures for GK-specific promotion. This procedure is commonly used in the TID program to evaluate GK's talent. Thus, it was deemed the most appropriate criterion variable to evaluate the assessment's validity.

Given previous findings (Gil et al. 2014), a potential influence in the coach rating based on GKs' age or anthropometric characteristics was checked. No significant differences between the talent groups regarding GKs' age, height, or weight were present ($0.149 \leq F\ (2, 117) \leq 1.341$; $p \geq .273$). Thus, a potential bias by these parameters can be precluded.

Statistical analysis

To evaluate the assessment's psychometric properties on a global level, a *score* that includes performance in each test with an equal weight was calculated for both T1 and T2. This score is represented by the mean out of all Z-transformed individual test performances.

Reliability (objective 1)

Following Mokkink et al. (2010b), three domains of reliability were evaluated to estimate the degree of measurement error in the data. These are represented by the assessment's internal consistency, the test–retest reliability ('relative reliability'), and the agreement ('absolute reliability'; Atkinson and Nevill 1998; de Vet et al., 2006).

To display the assessment's *internal consistency*, Cronbach's α for all individual tests at T1 and T2 was calculated (Tavakol and Dennick 2011). It was also evaluated if Cronbach's α differed if one test was deleted from the calculation.

To assess the *test–retest reliability*, Pearson's correlation coefficient (r_{xy}) between test performances at T1 and T2 was computed for the total sample.[2] As the sample's heterogeneity could affect the reliability (De Vet et al., 2006), partial correlations controlled for the age groups ($r_{xy\text{-}age}$) and the two regions of data collection ($r_{xy\text{-}region}$) were calculated. As these regions similarly represent the two included birth cohorts, no additional control for cohort effects was required. Reliability coefficients were displayed with their 95% confidence intervals. These were computed via bootstrapping for partial correlations using the *boot* package (Canty and Ripley 2024). A uniform guideline to interpret reliability coefficients of motor tests is not available as highly dependent on an assessment's practical purpose (Weir 2005). This is why rather conservative thresholds

used in medical contexts were applied that classify coefficients as 'poor to moderate' (0.5–0.75), 'good' (0.76–0.9), and 'acceptable for clinical measures' (>0.9; Portney and Watkins 2015).

To display the assessment's *agreement*, the standard error of measurement (SEM; Equation 1), smallest detectable change (SDC; Equation 2), and coefficient of variation (CV in %; Equation 3) were evaluated (de Vet et al., 2006; Mokkink et al. 2020). To calculate the SEM and SDC for each individual test item, a variable representing the performance difference between T1 and T2 was computed. From this variable, the standard deviation (SD_{diff}) was calculated. To compute the CV, the standard deviation ($SD_{T1,T2}$) and mean ($M_{T1,T2}$) were calculated from each individual's test performances at T1 and T2.

$$\mathrm{SEM} = \frac{SD_{diff}}{\sqrt{2}} \tag{1}$$

$$\mathrm{SDC} = 1.96 * \mathrm{SD}_{\mathrm{diff}} \tag{2}$$

$$\mathrm{CV}(\%) = \left(\frac{SD_{T1,T2}}{M_{T1,T2}}\right) * 100 \tag{3}$$

To display individual variations and systematic differences between test performances at T1 and T2 graphically, Bland Altman Plots were used with the mean of performances at T1 and T2 on the x-axis and the differences between these measurement points on the y-axis (Bland and Altman 1999). The mean difference and corresponding 95% limits of agreement (LoA) were computed and marked within the graphs according to Bland and Altman (2003) by using the *Blandr* package (Datta 2017). Additionally, paired t-tests were used to evaluate significant differences between performances at T1 and T2 (Atkinson and Nevill 1998). Cohen's *d* served as effect size (Cohen 1992) calculated as Cohen's d_{av} (Equation 10 in Lakens 2013).

To investigate the agreement for the score, raw values of individual tests at T2 were Z-transformed by subtracting the respective mean value and dividing by the standard deviation from T1. The score for T2 was calculated by the mean out of all these Z-transformed individual test performances.

Structural validity (objective 2)

Due to the inductive test development, exploratory factor analysis (EFA) with the Z-transformed data of all six tests at T1 was performed. While all tests were based upon motor coordination relating to ball control (i.e., motor tasks in a dynamic environment), this allows the identification of potentially underlying dimensions of the assessment. The number of factors was estimated by the Kaiser-Guttmann criterion suggesting to retain as many factors as there are eigenvalues ≥1 and a scree-test. Oblique promax-rotation for all retained factors served to present the final solution (Thompson 2004).

To cross-validate the assessment's factorial structure identified at T1, confirmatory factor analysis (CFA) with the data collected at T2 was performed. In these analyses, fixed factor loadings were defined based on the EFA results. As model-fit indicators, the χ2-value, root mean square error of approximation (*RMSEA*), and the standardized root mean square residual (*SRMR*), as well as the comparative fit index (*CFI*) and the Tucker – Lewis Index (*TLI*) were considered. Acceptable model-fit was assumed if the corrected χ^2 -squared values (i.e., divided by the tests degrees of freedom) are ≤3.0, *RMSEA* and *SRMR* ≤ .08, and *CFI*/*TLI* > .90 (Hu and Bentler 1999; Marsh et al. 2004).

Concurrent validity (objective 3)

To display the assessment's *concurrent validity* – as one domain of criterion validity (Mokkink et al. 2010a) – the diagnostic accuracy was evaluated with Receiver Operating Characteristic (ROC) Curves at both T1 and T2 (Mandrekar 2010). These analyses were conducted regarding two different practical purposes of a talent assessment: First, the identification of GKs with sufficient talent to get promoted at one of the DFB BCs. Therefore, the rating was dichotomized to assess the accuracy in identifying GKs considered worth promoting (i.e., groups A and B) in relation to detecting the least talented GKs (i.e., group C).

The second practical purpose is the identification of 'elite youth GKs'. Thus, the accuracy in identifying A-rated (i.e., group A) GKs in relation to detecting lower-rated individuals (i.e., group B and C) was investigated. It should be noted that dichotomization removes some of the data's variance but given the fact that this is how such assessments may be used in practice, such evaluation seems relevant. To control for the age groups and regions of data collection, covariate-adjusted ROCs (AROCs) and the respective adjusted area under the curve (*AAUC*) were calculated using the AROC package (Machado e Costa and Braga 2020). A significant deviation of the ROC curve from the diagonal (i.e., 45-degree line) and the corresponding AAUC was checked by considering the curves' 95% confidence intervals.

More detailed insights on the concurrent validity were evaluated with analyses of covariance (ANCOVAs) considering the test performance as dependent variable and categorical expert ratings of a GK's talent (i.e., groups A, B, or C) as independent variable. This shows which talent groups differ in their test performance and further allow for comparison of the individual test's diagnostic power. Partial eta squared (η_p^2) served as the effect size to display the assessment's concurrent validity at T1 and T2. Computing ANCOVAs by controlling for both the age groups and the regions of data collection as covariates was chosen given comparably small sample sizes in each age and talent group (Hecksteden et al. 2022). To check the independence of the covariates and the coach rating as the independent variable, chi-squared tests were performed, revealing no significant differences for neither the age groups (χ^2 (6) = 5.88, p = .44) nor regions of data collection (χ^2 (2) = 3.61, p = .17).

The test power is a critical issue in talent research in general (Bergkamp et al. 2019) and specifically in GK-specific talent research (Vahia and Kelly 2024). Therefore, sensitivity analyses were performed to determine the size of a possibly detectable effect utilizing G*Power (version 3.1.9.7; Faul et al. 2009) with the predetermined parameters for ANCOVAs ($\alpha = 0.05$, $1 - \beta = 0.80$, $N = 120$ across three groups, and two covariates). The analyses were sensitive enough to detect amoderate effect size (i.e., $\eta_p^2 = .076$).

Group-based differences were tested for significance via contrasts. To confirm the criterion validity, it was hypothesized that higher-rated GKs outperformed lower-rated GKs (i.e., A > C,

A > B, B > C). Based on the covariate-adjusted descriptive statistics, differences between talent groups were displayed as Cohen's *d* (computed as the mean difference divided by the pooled standard deviation; Cohen 1992). Effect sizes were provided irrespective of their significance, considering the notable differences in sample sizes across the three groups.

Statistical analyses relating to the agreement, EFAs, and ANCOVAs were processed utilizing IBM SPSS (version 28). CFAs were performed in Mplus (version 8.2; Muthén and Muthén 2017). Test–retest reliability, Bland Altman, and AROC analyses as well as the data visualization using the ggplot2 package (Wickham 2016) were computed in R (version 4.2.2; R Core Team 2021). The alpha level was set at 5% in all analyses.

Results

Reliability (objective 1)

The test battery demonstrates good *internal consistency* at T1 (α = .78) and T2 (α = .78). At both measurement points, Cronbach's α differs only marginally when one test is deleted from the calculation (Δα ≤ .06).

The *test–retest reliability* coefficients are presented in Table 3. The score shows 'good' test–retest reliability for the total sample (r = .90). When controlling for age groups, the score's reliability is lower, but still 'good' (r = .80). The individual tests show 'good' (i.e., JBC, BWR II) and 'poor to moderate' reliability (i.e., DBB, BWR I, TSC, BTC). When controlling for age groups, the reliability of the JBC is still 'good', the further tests reach 'poor to moderate' coefficients. These age-adjusted coefficients are slightly lower in the JBC, TSC, and BTC (.02 ≤ Δr ≤ .06). More pronounced differences compared to the age-unadjusted coefficients are found for the DBB, BWR I, and BWR II (.13 ≤ Δr ≤ .22). Partial correlations controlled for the region where the data was collected are in some test marginally lower compared to the unadjusted coefficients (.00 ≤ Δr ≤ .08).

Results on the *agreement* across T1 and T2 are displayed in Table 3 and Figure 1. The score is affected by substantial measurement error (SEM = 2.3; SDC = 4.5; CV = 3.77%) and significantly better test performances at T2 are present (d = 0.75). All test items show substantial, yet differently pronounced measurement errors (1.3 ≤ SEM ≤ 14.7, 3.5 ≤ SDC ≤ 40.8; 7% ≤ CV ≤ 45%). The DBB reveals the most pronounced measurement error. These differences between the test performances at T1 and T2 are also visible in Bland Altman analyses and plots (Figure 1). These further illustrate significantly better test performances at T2 across all individual tests with small to moderate effect sizes (0.16 ≤ d ≤ 0.69). The TSC and BTC demonstrate the least pronounced systematic improvements, but outliers are similarly present.

Structural validity (objective 2)

A significant Bartlett test (χ^2 (15) = 189.22, p < .001) and – as indicated by the good internal consistency (α_{T1} = .78) – Kaiser–Meyer–Olkin criterion of .80 reveal a sufficient correlation between the six individual tests at T1 to perform an EFA. Kaiser's criterion and the scree plot (Supplement II) yield an empirical justification for a one-factor solution with an eigenvalue of 2.92 and 49% of explained variance.[3] The pattern matrix shows the highest factor loadings for DBB, BWR I, and BWR II (.67 ≤ λ ≤ .79) and, thus, communalities between 45% and 62% of explained item variance (Table 4). The three further tests reveal lower, yet still acceptable loadings (λ ≥ .47) with communalities from 22% to 46%.

To cross-validate the one-factor solution proposed by the EFA at T1, a CFA was conducted using the data collected at T2. The cross-validation shows insufficient model-fit with a significant χ^2 -test ($\chi^2(9)$ = 5.53, p < .001) and poor model-fit indices *RMSEA* [90%-CI] = .19 [.14; .25], *CFI* = .79, *TLI* = .68, *SRMR* = .09). Thus, the proposed one-factor solution could not be replicated at T2. Yet, all tests reveal substantial factor loadings (.54 ≤ λ ≤ .82; all p < .001), except for BTC with a lower, though still significant factor loading (λ = .34 p < .001). Consistent with the factor loadings found in the EFA, DBB, BWR I, and BWR II show the highest factor loadings (.62 ≤ λ ≤ .82).

Table 3. Results on the test–retest reliability ('relative reliability') and agreement ('absolute reliability').

		Test-Retest Reliability			Agreement			
	N	r_{xy} [95% CI]	$r_{xy\ -\ Age\ Group}$ [95% CI]	$r_{xy\ -\ Region}$ [95% CI]	SEM [points]	SDC [points]	CV [%]	SMD [Cohen's *d*]
Score	120	.90*** [.86; .93]	.80*** [.71; .87]	.90*** [.85; .93]	2.3	4.5	3.77 [3.42; 4.13]	0.75*** [0.38; 1.12]
JBC	120	.80*** [.73; .86]	.75*** [.65; .83]	.75*** [.64; .82]	3.5	9.7	15.92 [13.47; 18.37]	0.56*** [0.20; 0.93]
DBB	120	.65*** [.54; .75]	.52*** [.37; .65]	.65*** [.54; .76]	14.7	40.8	44.73 [38.81; 50.66]	0.62*** [0.25; 0.99]
BWR I	120	.71*** [.61; .79]	.49*** [.24; .70]	.73*** [.55; .84]	2.5	6.9	7.27 [6.09; 8.44]	0.69*** [0.32; 1.06]
BWR II	120	.78*** [.70; .84]	.64*** [.53; .74]	.77*** [.70; .84]	1.5	4.1	11.83 [10.23; 13.43]	0.53*** [0.16; 0.89]
TSC	120	.73*** [.63; .80]	.67*** [.55; .77]	.72*** [.60; .80]	1.9	5.2	31.67 [25.38; 37.95]	0.16* [−0.20; 0.55]
BTC	120	.71*** [.61; .79]	.69*** [.60; .78]	.63**** [.50; .74]	1.3	3.5	18.64 [15.45; 21.83]	0.33*** [−0.03; 0.87]

Note. ***p < 0.001; JBC = Juggling-Ball-Change; DBB = Double-Ball-Bouncing; BWR I = Ball-Wall-Rebounder – one ball; BWR II = Ball-Wall-Rebounder – two balls; TSC = Throwing-Sitting-Catching; BTC = Bouncing-Turning-Catching; r_{xy} = product-moment correlation coefficient (Pearson); $r_{xy\ -Age\ Group}$ = Partial correlation controlled for the age groups U12 to U15; $r_{xy\ -\ Region}$ = Partial correlation controlled for the two regions (i.e., Westphalia & Württemberg) in which the data was collected; SDC = Smallest Detectable Change; SEM = Standard Error of Measurement; SMD = Standardized Mean Difference; CV = Coefficient of Variation.

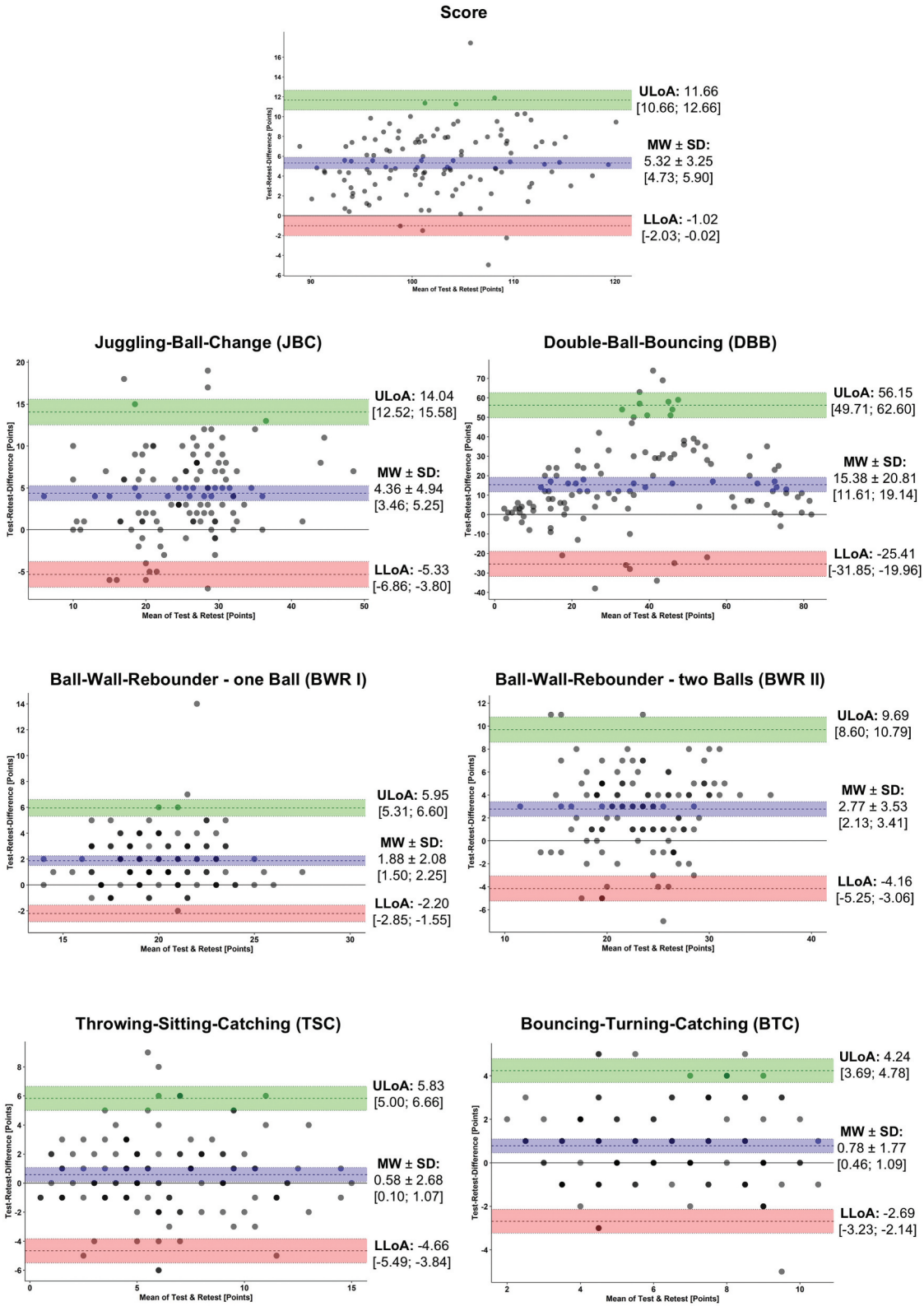

Figure 1. Bland Altman plots to display differences in performance at the first (T1) and second measurement point (T2) for the score and individual test items. ***Note.*** Green (upper limit of agreement; ULoA) and red (lower limit of agreement; LLoA) lines, as well as their shadings, represent the 95% confidence intervals.

Concurrent validity (objective 3)

The overall test performance, as reflected in the score, reveal *concurrent validity* for two practical purposes. Regarding the identification of all talented GKs (i.e., groups A and B; sensitivity) in relation to the identification of C-rated GKs (i.e., specificity), AROC analyses demonstrate significant and nearly equivalent diagnostic accuracy at T1 (*AAUC* = .68 [.56; .79]; Figure 2(a) and T2 (*AAUC* = .67 [.56; .78]; Figure 2(b). This

Table 4. Results of the exploratory factor analysis with the data collected at T1.

		Factor Loadings							
	Factors and communality	JBC	DBB	BWR I	BWR II	TSC	BTC	Eigenvalue	Explained variance
1-Factor solution	Factor 1	.59	.67	.79	.68	.47	.50	2.916	48.61%
	h^2	.34	.45	.62	.46	.22	.25		

Note. JBC = Juggling-Ball-Change; DBB = Double-Ball-Bouncing; BWR I = Ball-Wall-Rebounder – one ball; BWR II = Ball-Wall-Rebounder – two balls; TSC = Throwing-Sitting-Catching; BTC = Bouncing-Turning-Catching.

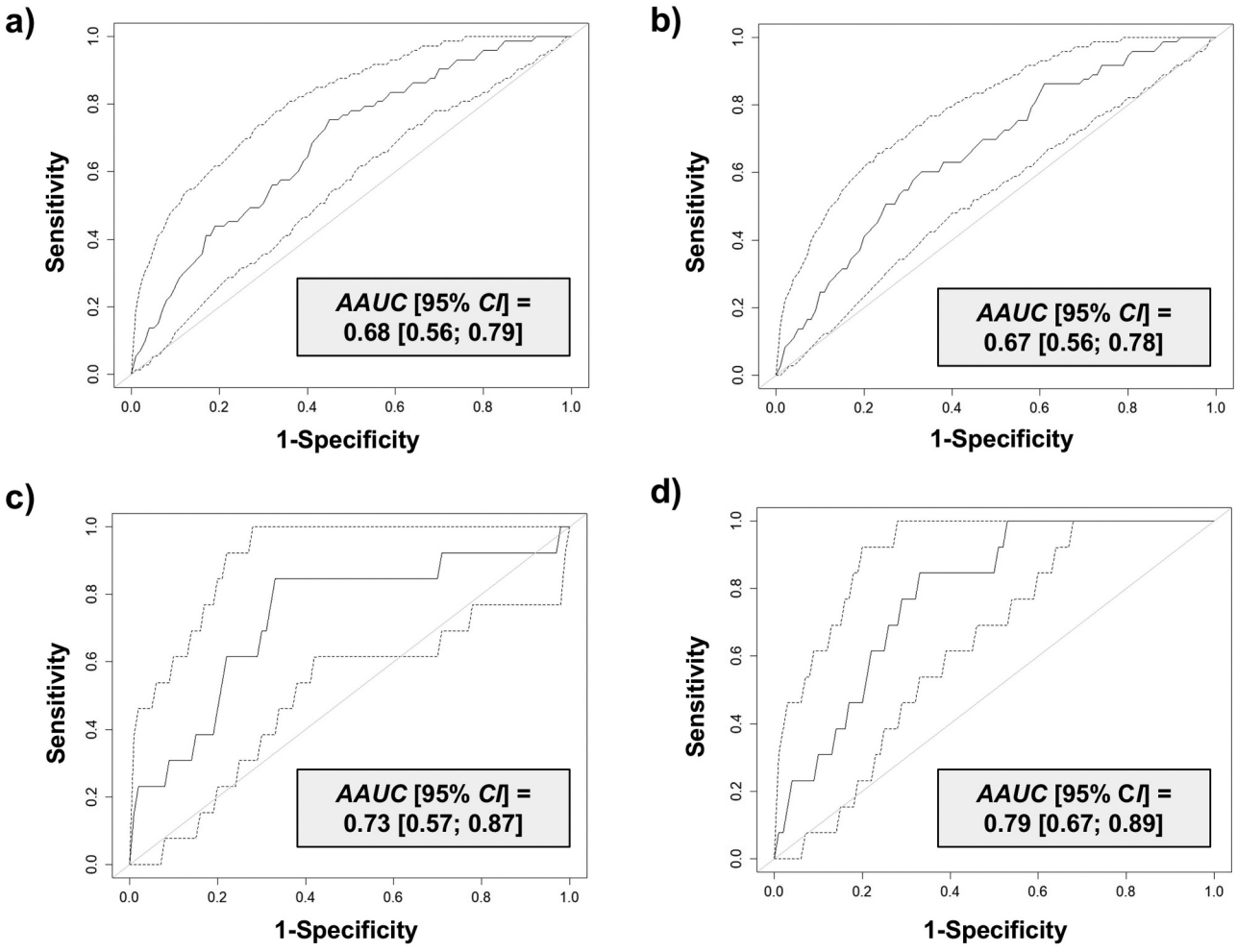

Figure 2. AROC curves displaying the diagnostic accuracy of the score (i.e., the overall test performance). ***Note.*** The analyses were guided by two potential practical purposes of a talent assessment. The first purpose (Figures 2(a,b) for T1 and T2, respectively) is the identification of GKs who are worth promoting in the TID program (i.e., groups A and B) in relation to detecting the least talented GKs (i.e., group C). The second practical purpose (Figures 2(c,d) for T1 and T2, respectively) is the identification of 'elite youth GKs' (i.e., group A) in relation to detecting lower-rated individuals (i.e., group B and C).

indicates a 68% (T1) and 67% (T2) chance of correct assignment based on the score.

Regarding the identification of elite youth GKs (i.e., group A; sensitivity) in relation to lower-rated individuals (i.e., groups B and C; specificity), significant diagnostic accuracy is found at T1 (*AAUC* = .73 [.57; .87]; Figure 2(c) and T2 (*AAUC* = .79 [.68; .89]; Figure 2(d). This indicates a 73% (T1) and 79% (T2) chance of correct assignment based on the score.

These results generally confirm that the assessment reveals some diagnostic power. More detailed insights on the concurrent validity were evaluated with ANCOVAs by considering the test performance as the dependent variable and the talent groups as the independent variable. Regarding the score, these analyses show significant differences between the talent groups with a large effect size at T1 ($F(2, 115) = 10.30$, $p < .001$, $\eta_p^2 = .15$) and with an even slightly higher effect size at T2 ($F(2, 115) = 11.83$, $p < .001$, $\eta_p^2 = .17$). Subsequently performed contrasts to compare the test performances across talent groups show significantly better performances of higher rated GKs, thus, supporting the assessment's concurrent validity. The contrasts to compare the talent groups show the largest differences between A-rated and C-rated GKs ($d_{T1} = 1.36$; $d_{T2} = 1.47$).

Table 5. Results of the ANCOVAs controlled for age groups and regions in which the data were collected and contrasts to compare test performances between three talent groups at the second measurement point (T2).

		$M_{adjusted} \pm SE$ 95% CI [LL; UL]			ANCOVA			Contrasts Cohen's *d* 95% CI [LL; UL]		
Test	*N*	A (*n* = 13)	B (*n* = 60)	C (*n* = 47)	*F*(2, 115)	*p*	η_p^2	A > C	A > B	B > C
Score	120	104.34 ± 1.23 [101.90; 106.78]	100.36 ± 0.57 [99.22; 101.49]	97.77 ± 0.65 [96.48; 99.07]	11.83	<.001	**.17*** **	**1.47*** ** **[0.80; 2.14]**	**0.90** ** **[0.28; 1.52]**	**0.58** ** **[0.19; 0.97]**
JBC	120	30.00 ± 1.68 [26.69; 33.33]	28.08 ± 0.78 [26.54; 29.63]	25.48 ± 0.89 [23.73; 27.24]	3.80	.025	**.06***	**0.74*** **[0.11; 1.37]**	0.32 [−0.28; 0.92]	**0.43*** **[0.04; 0.82]**
DBB	120	57.30 ± 6.18 [45.07; 69.54]	45.04 ± 2.87 [39.35 50.73]	40.98 ± 3.27 [34.49; 47.46]	2.70	.072	.05	**0.73*** **[0.10; 1.36]**	0.55 [−0.01; 1.16]	0.18 [−0.20; 0.57]
BWR I	120	22.27 ± 0.54 [21.19; 23.34]	21.04 ± 0.25 [20.54; 21.54]	20.51 ± 0.29 [19.94; 21.08]	4.06	.020	**.07** **	**0.89** ** **[0.25; 1.52]**	**0.63*** **[0.03; 1.24]**	0.27 [−0.11; 0.65]
BWR II	120	26.10 ± 1.16 [23.80; 28.39]	25.17 ± 0.54 [24.10; 26.23]	23.38 ± 0.61 [22.16; 24.59]	3.28	.041	**.05***	**0.65*** **[0.03; 1.28]**	0.22 [−0.38; 0.82]	**0.43*** **[0.04; 0.81]**
TSC	120	9.46 ± 0.87 [7.73; 11.19]	6.85 ± 0.41 [6.05; 7.66]	5.57 ± 0.46 [4.66; 6.49]	7.87	<.001	**.12*** **	**1.23*** ** **[0.58; 1.89]**	**0.82** ** **[0.21; 1.44]**	**0.40*** **[0.02; 0.79]**
BTC	120	7.45 ± 0.51 [6.42; 8.47]	6.61 ± 0.24 [6.13; 7.09]	6.12 ± 0.27 [5.59; 6.66]	2.68	.073	.04	**0.72*** **[0.09; 1.35]**	0.45 [−0.15; 1.06]	0.26 [−0.20; 0.65]

Note. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; JBC = Juggling-Ball-Change; DBB = Double-Ball-Bouncing; BWR I = Ball-Wall-Rebounder – one ball; BWR II = Ball-Wall-Rebounder – two balls; TSC = Throwing-Sitting-Catching; BTC = Bouncing-Turning-Catching; The age groups (i.e., U12-U15) and the region in which the data was collected (i.e., Westphalia & Württemberg) were included as covariates in the ANCOVAs.

Lower, but still large effect sizes are found for comparisons of A-rated and B-rated GKs ($d_{T1} = 0.81$; $d_{T2} = 0.90$). Moderate effect sizes are present for comparisons of B-rated and C-rated GKs ($d_{T1} = 0.56$; $d_{T2} = 0.58$). These results confirm the assessment's concurrent validity at both T1 and T2. Yet, the overall slightly higher diagnostic power at T2 suggests better considering data from the second measurement point.

For this second measurement point, findings regarding the concurrent validity of all six individual tests are displayed in Table 5. In four out of six individual tests, the ANCOVAs show

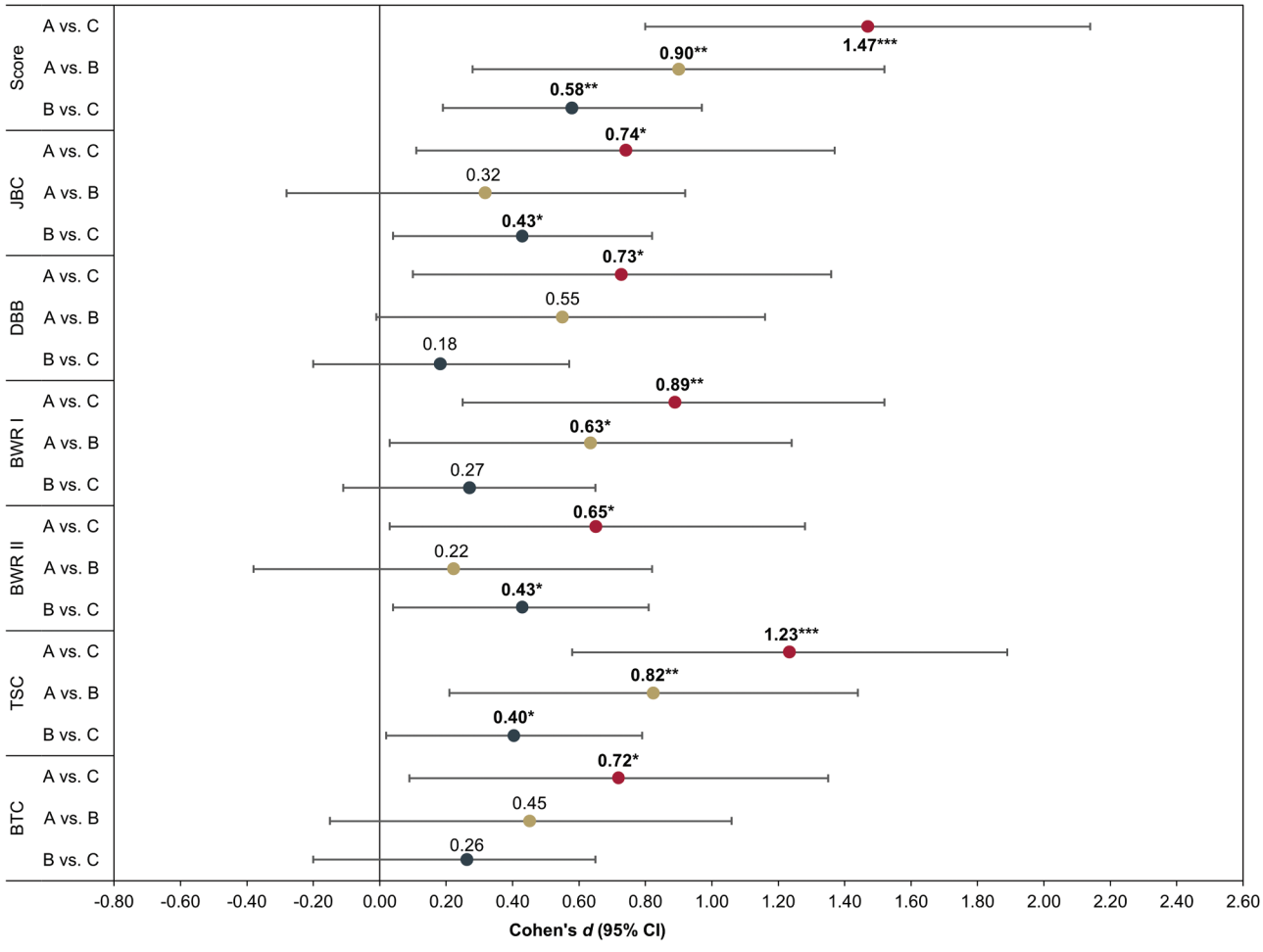

Figure 3. Standardized mean differences to display differences in performance of GKs assigned to three talent groups at T2. ***Note.*** A positive effect size represents better performance of higher-rated GKs. Error bars display 95% confidence intervals.

significant differences between the groups with moderate effect sizes (.05 ≤ η_p^2 ≤ .12). Partial eta-squared of the two non-significant models are still moderate but it should be noted that the effect sizes are below the threshold for a detectable effect size when considering the analyses' sensitivity. Across all individual tests, the elite GKs (i.e., group A) significantly outperform the least talented GKs (i.e., group C) with moderate to large effect sizes (0.65 ≤ *d* ≤ 1.23; Figure 3). The comparisons of GKs in groups A and B show two significant differences with moderate to large effect sizes (0.63 ≤ *d* ≤ 0.82). Two non-significant differences with moderate effect sizes are found (0.45 ≤ *d* ≤ 0.55) and two further non-significant differences show small effect sizes (0.22 ≤ *d* ≤ 0.32). The comparisons between B-rated and C-rated GKs show significant differences with small effect sizes in three tests (0.40 ≤ *d* ≤ 0.43), while the remaining three tests reveal no significant differences with small effect sizes (0.18 ≤ *d* ≤ 0.27).

Discussion

A GK-specific motor coordination assessment aiming to support TID in football was co-designed by researchers and practitioners assembled into an expert panel. Ratings by experts not involved in the test development demonstrate the test battery's content and face validity, as well as the practical feasibility. Findings on the assessment's psychometric properties reliability (objective 1) and validity (objectives 2 & 3) are ambivalent. The score representing the GKs' overall test performance shows good reliability. Furthermore, the score reveals a practically relevant diagnostic power in identifying the more talented GKs, thus demonstrating concurrent validity. Therefore, corresponding to previous talent research using motor coordination tests across sports (e.g., Vandorpe et al. 2012; O'Brien-Smith et al. 2019), the test battery offers potential to support the identification of goalkeeping talent. Yet, the findings also underline that the assessment's diagnostic accuracy is not sufficient to justify decisions on an individual level (e.g., [de]selections). In addition, the test–retest reliability of some tests is only poor to moderate, and substantial measurement error affects the agreement across repeated measurements. Furthermore, the exploration of the structural validity indicates that the test battery does not represent a uniform construct. Therefore, the findings also uncover different areas where the assessment should be optimized for enhancing its potential to support GKs' talent identification in practice.

Comparing findings on the *reliability* with other studies is difficult due to several factors: i) longer test–retest periods, ii) heterogeneous study samples without the control for covariates (e.g., age, sex), iii) unknown measurement errors, or iv) the absence of reliability coefficients. A comparison of the test–retest reliability for the total sample and a study with table tennis players (7–12 yrs.) is possible that report slightly higher coefficients of an 'eye-hand coordination test' between attempts performed on the same day (*ICC* = 0.83; Faber et al. 2014). Visual analysis indicates less systematic improvements between test attempts, but random errors and outliers are similarly present. Further studies evaluating comparable tests show heterogeneous test–retest reliability coefficients (.31 ≤ *ICC* ≤ .91; Faber et al. 2015), and some tests similarly reveal limited agreement (3% ≤ CV ≤ 43%, Faber et al. 2015; 10% ≤ CV ≤ 12%, Matthys et al., 2013). Therefore, the presented findings widely correspond with previous research regarding challenges in achieving proper test–retest reliability and especially agreement in ball-related coordination tests.

Considering that the assessment is mainly intended to aid discriminative purposes (i.e., identifying talented individuals among a group of GKs), 'relative reliability' (e.g., test-retest-reliability) is principally more relevant than 'absolute reliability' (e.g., agreement; de Vet et al., 2006). Nevertheless, if the assessment aims to systematically support talent identification in a large-scale TID program, reducing measurement error is essential to establish reference values for comparing GKs across different regions and birth cohorts. Overcoming this issue may relate to optimizing the testing guidelines. Similar to previous studies (Faber et al. 2015; Matthys et al., 2013; Höner et al., 2015), data close to participants' *baseline performance* (i.e., the initial performance on a given task; Baltes 1987) was analyzed given only short familiarization before the actual testing at T1. Yet, ongoing familiarization across attempts led to considerable improvements from T1 to T2 (0.16 ≤ *d* ≤ 0.69). Thus, it is worthwhile to explore how many attempts are required until a plateau in performance is reached, to potentially consider data with less measurement error, which is also expected to improve the test's reliability (Weir 2005).

Regarding the assessment's *structural validity*, a unidimensional structure of the assessment as suggested by the EFA would be coherent with other assessments because all tests assess motor coordination related to ball control (i.e., based on motor tasks in a dynamic environment). In previous research, ball-related motor tests often load on a latent factor termed 'object/ball control' or similar (e.g., Faber et al. 2015; Herrmann and Seelig 2017; Garn and Webster 2021). However, while these and further assessments include comparable tests to those evaluated in the present study, their underlying constructs often differ (e.g., 'motor skills', 'fundamental motor skills'; 'motor competence'). This also reflects the variety of taxonomies and theories to classify motor skills and abilities (Lämmle et al. 2010; Hands et al. 2018; Magill and Anderson 2021), as well as concepts aiming to collectively assess motor performance dispositions as general or basic motor competencies (Fransen et al. 2014; Herrmann and Seelig 2017). In contrast to these studies, the presented assessment was guided by the assumption that high levels of abilities relevant to a certain field (i.e., goalkeeping) are one of many indicators of a person's potential to develop GK-specific performance attributes. Guided by the GK literature (e.g., Busch 2017) and informed by the expertise of practitioners from the expert panel that developed the assessment, it is assumed that this can be evaluated through purposefully designed motor coordination tests focusing on ball control.

Considering the factor loadings of individual tests at T1, the TSC and BTC exhibit lower but still sufficient loadings. This could be explained by more pronounced gross-motor demands (i.e., sitting down, standing up, jumping; turning around) compared to the remaining tests. On a confirmatory level, the TSC again shows a lower factor loading (λ = .54, $p < .001$), but especially the BTC (λ = .34, *p*

< .001) seemed to not fit optimally in the test battery. One reason for the poor model fit found for the replication of the unidimensional structure at T2 can be differently pronounced measurement errors across individual tests. In this regard, it seems worth noting that the TSC and BTC demonstrate the least systematic improvements from T1 to T2. Again, these findings hint at a potential contribution from optimized testing guidelines aiming to reduce measurement errors that eventually affect the assessment's structural validity.

Given these findings, the score used as a global estimate of the test performance therefore does not represent a uniform theoretical construct. However, this score as a proxy for subjectively rated goalkeeping talent (i.e., the criterion), shows higher diagnostic power in classifying the GKs compared to individual tests, thereby enhancing the *concurrent validity*. This finding aligns with previous talent research which uses scores computed from different (heterogeneous) tests to improve the diagnostic and predictive power (e.g., Höner et al. 2015).

The highest diagnostic accuracy in identifying A-rated GKs and moderate to large group-based mean differences between A- and C-rated GKs indicate that the assessment is more suitable for identifying the most talented individuals. In addition, the somewhat higher effect sizes found at T2 reveal associations with the paradigm of *testing the limits* used in cognitive diagnostics (Baltes et al. 2007). Specifically, more talented GKs outperformed less talented individuals to a greater extent when considering test data closer to a GK's baseline reserve capacity (i.e., the upper range of an individual's performance potential at a given time point, Baltes 1987). Yet, the available data does not allow for clear conclusions on how close GKs were to their baseline reserve capacity. Therefore, in line with strategies to potentially improve the assessment's test–retest reliability and agreement, a longer time for familiarization, more test attempts, or even practice seem worth evaluating in the future.

Adjusting the testing guidelines toward more attempts is, however, not trivial as other factors could negatively affect the repeated motor test performance (e.g., fatigue, decreasing concentration), or the practical feasibility (e.g., time needed). Therefore, respective adjustments could make it necessary to reduce the overall number of tests. This decision can be facilitated by comparing the effect sizes to identify tests with the highest concurrent validity. In this evaluation and the pilot study (Bergmann et al. 2021), the TSC shows the highest diagnostic power across all talent groups. On a group-based level, there are significant performance differences in the BWR I between A-rated GKs compared to lower-rated groups, but not between B- and C-rated GKs. The JBC and BWR II discriminate A- and B-rated GKs from group C with almost similar effect sizes. The DBB and BTC fail to reach significance on a group-based level, but the effect sizes are below the threshold of the analyses' sensitivity. As far as adjustments in the testing guidelines do not improve these tests' concurrent validity, if the DBB remains substantially affected by measurement error, and if the BTC still shows a considerably lower factor loading compared to the remaining tests, eliminating these tests from the test battery should be contemplated.

To summarize, this study supports that the GK-specific motor coordination assessment reveals the potential to support talent identification in football. However, the findings also hit at risks if the assessment's psychometric weaknesses and the resulting limitations for practical use are disregarded. Like other assessments established in TID programs across sports (e.g., Vandorpe et al. 2012; Faber et al. 2015; Matthys at al., 2013; Sieghartsleitner, Zuber, Zibung, Conzelmann 2019), the limited accuracy in assigning talented individuals underlines that such assessments can only supplement and not substitute experts' subjective judgments. Yet, if the measurement error can be reduced, reference values would allow for comparison of GKs in the nationwide TID program, which represents a valuable contribution to enrich the coach's eye.

Limitations

Although the sample of youth GKs assessed in this study is large compared to other studies, the sensitivity of analyses was still limited. This limitation is particularly notable due to the preselected sample, where small to moderate effect sizes (i.e., below the study's sensitivity) could still be relevant. Additionally, one somewhat expected consequence of the short test–retest period is considerable performance improvements from T1 to T2. Achieving a longer, yet similar test–retest period across GKs was not feasible due to the specific target group so this rather conservative approach was preferred. Furthermore, defining thresholds for coefficients related tomeasurement error (e.g., CV, LoAs) is difficult due to lacking studies using comparable assessments that further evaluated the amount of 'acceptable' error while still ensuring practically relevant validity. Additionally, ratings from GK-coaches were used to operationalize a GK's talent as the criterion variable for evaluating the concurrent validity of the assessment. This was chosen because experts' judgments are established and often regarded as the most relevant approach for categorizing GKs in terms of their talent in practice. However, it is important to note that objective assessments are intended to enhance and supplement these subjective evaluations (e.g., Höner et al., in revision; Sieghartsleitner, Zuber, Zibung, Conzelmann 2019), suggesting that reliance solely on expert ratings may not yield the most accurate method for identifying talent. Lastly, all analyses needed to be conducted across four age groups. While the analyses on the reliability and concurrent validity were controlled for covariates, the exploration of structural validity was based on this age-heterogeneous sample.

Future perspectives

Optimizing the assessment's test–retest reliability and agreement similarly like the concurrent validity may be achieved by considering data closer to a GK's *baseline reserve capacity* (Weir 2005; Baltes et al. 2007). Therefore, future investigations on the time point when a plateau in the test performance occurs in different age and talent groups are required to potentially optimize the testing guidelines. If these modifications do not contribute to some test's psychometric properties, their elimination should be considered. In addition, the *test's purposeful weighting* (e.g., based on discriminant analyses) could further improve the score's diagnostic accuracy regarding different practical purposes. After a few years, the present data will allow for evaluation of the test battery's *predictive validity* which is one of the most relevant requirements to inform TID.

As established and prognostically relevant objective and GK-specific subjective assessments were already used in the DFB TID program (Höner et al., in revision), the *incremental validity* of this new assessment needs to be analyzed.

Finally, considering different processes that shape the talent pathway (Williams et al. 2020), this study evaluates the assessment's potential to inform the identification of talented individuals within a pool of GKs already participating in a TID program. Given the focus on motor coordination as one of many indicators of a GK's giftedness (Gagné 2021), this assessment could eventually inform the *talent detection* of subjects not yet acting as a GK. This could address individuals already participating as outfielders as well as children not yet playing football.

Notes

1. Preckel et al. (2020) use the term *aptitude* that relates to, for example, general and domain-specific abilities.
2. The COSMIN guidelines recommend the *ICC* to evaluate the reliability. The *ICC* in a two-way random model (type consistency) leads to the same results as Pearson's *r*, if two measurements were considered. Thus, Pearson's *r* was preferred as it allows to control for covariates due to the sample's heterogeneity.
3. While these findings support a one-factor solution, it should be noted that a two-factor solution reveals an eigenvalue of .989 and 65% explained variance. Thus, the recommendation for a two-factor solution was just below the defined threshold of 1.0.

Acknowledgements

We would like to thank the staff of the DFB's Department for Talent Development for productive discussions in several meetings.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study is part of the research project 'Scientific support of the DFB's Talent Promotion Program', which is funded by the German Football Association (Deutscher Fußball-Bund, DFB).

Notes on contributors

Fynn Bergmann: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – First Draft.

Florian Schultz: Writing – Review & Editing.

Job Fransen: Writing – Review & Editing.

Oliver Höner: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing.

ORCID

Fynn Bergmann http://orcid.org/0000-0001-8323-3157
Job Fransen http://orcid.org/0000-0003-3355-1848
Oliver Höner http://orcid.org/0000-0002-3108-1531

Data availability statement

The dataset, codes, and outputs (SPSS, Mplus, and R) are publicly available in the open science framework: https://osf.io/exuhq.

Open scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://osf.io/exuhq

References

Aadland KN, Nilsen AKO, Lervåg AO, Aadland E. 2022. Structural validity of a test battery for assessment of fundamental movement skills in Norwegian 3-6-year-old children. J Sports Sci. 40(15):1688–1699. doi: 10.1080/02640414.2022.2100622.

Atkinson G, Nevill AM. 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 26 (4):217–238. doi: 10.2165/00007256-199826040-00002.

Baker J, Wilson S, Johnston K, Dehghansai N, Koenigsberg A, de Vegt S, Wattie N. 2020. Talent research in sport 1990–2018: a scoping review. Front Phychol. 11:607710. doi: 10.3389/fpsyg.2020.607710.

Baltes PB. 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Dev Phychol. 23 (5):611–626. doi: 10.1037/0012-1649.23.5.611.

Baltes P, Lindenberger U, Staudinger U, Lerner RM, Lerner RM. 2007. Life span theory in developmental psychology. In: Damon W, R. M. Lerner, & R. M. Lerner Eds Handbook of child psychology. 569–664. 10.1002/9780470147658.chpsy0111.

Bergkamp T, Niessen A, den Hartigh R, Frencken W, Meijer R. 2019. Methodological issues in soccer talent identification research. Sports Med. 49(9):1317–1335. doi: 10.1007/s40279-019-01113-w.

Bergmann F, Danner J, Schultz F, Dugandzic D, Kompodietas E, Höner O. 2024. In: Testmanual: Ballgebundener Koordinationstest zur Talentdiagnostik von Fußballtorhütern in der frühen Adoleszenz (U11-U15). [Test manual. Ball-related motor coordination assessment for football goalkeepers during early adolescence (U11-U15)] 2nd Edition. Tübingen: Eberhard Karls Universität. https://uni-tuebingen.de/en/218829.

Bergmann F, Danner J, Schultz F, Pollmann D, Höner O. 2021. Talentdiagnostik von Torhütern-Entwicklung und Evaluation eines sensomotorischen Koordinationstests [Talent assessment in football goalkeepers. Development and evaluation of a motor coordination assessment]. Leistungssport. 51(1):50–55.

Bland JM, Altman DG. 1999. Measuring agreement in method comparison studies. Stat Methods Med Res. 8(2):135–160. doi: 10.1177/096228029900800204.

Bland JM, Altman DG. 2003. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecology. 22(1):85–93. doi: 10.1002/uog.122.

Busch C. 2017. Koordinationstraining für den Fußballtorwart [Coordination Practice for Football Goalkeepers]. Sportverlag Strauß.

Canty A, Ripley BD. 2024. Boot: bootstrap R (S-Plus) functions. R package version 1.3–30.

Cohen J. 1992. A power primer. Psychological Bull. 112(1):155. doi: 10.1037/0033-2909.112.1.155.

Coppens E, Laureys F, Mostaert M, D'Hondt E, Deconinck FJA, Lenoir M. 2021. Validation of a motor competence assessment tool for children and adolescents (KTK3+) with normative values for 6- to 19-year-olds. Front Physiol. 12:652952. doi: 10.3389/fphys.2021.652952.

Datta D. 2017. Blandr: a Bland-Altman method comparison package for R. doi: 10.5281/zenodo.824514.

Deprez D, Fransen J, Lenoir M, Philippaerts R, Vaeyens R. 2014. A retrospective study on anthropometrical, physical fitness, and motor coordination characteristics that influence dropout, contract status, and first-team playing Time in high-level soccer players aged eight to

eighteen years. J Strength Cond Res. 29(6):1692–1704. doi: 10.1519/JSC.0000000000000806.

de Vet HC, Terwee CB, Knol DL, Bouter LM. 2006. When to use agreement versus reliability measures. J Clin Epidemiol. 59(10):1033–1039. doi: 10.1016/j.jclinepi.2005.10.015.

Dugdale J, Sanders D, Myers T, Williams A, Hunter A. 2020. A case study comparison of objective and subjective evaluation methods of physical qualities in youth soccer players. J Sports Sci. 38(11–12):1304–1312. doi: 10.1080/02640414.2020.1766177.

Elleray A. 2021. Scientific approaches to goalkeeping in football. Staffordshire: Bennion Kearny Limited.

Faber IR, Elferink-Gemser MT, Faber NR, Oosterveld FGJ, Nijhuis-Van der Sanden MWG, Sampaio J. 2016. Can perceptuo-motor skills assessment outcomes in young table tennis players (7–11 years) predict future competition participation and performance? An observational prospective study. PLOS ONE. 11(2):e0149037. doi: 10.1371/journal.pone.0149037.

Faber IR, Koopmann T, Schipper-van Veldhoven N, Twisk J, Pion J, Oliveira RFS. 2023. Can perceptuo-motor skills outcomes predict future competition participation/drop-out and competition performance in youth table tennis players? A 9-year follow-up study. PLOS ONE. 18(2):e0281731. doi: 10.1371/journal.pone.0281731.

Faber IR, Nijhuis-Van Der Sanden MW, Elferink-Gemser MT, Oosterveld FG. 2015. The Dutch motor skills assessment as tool for talent development in table tennis: a reproducibility and validity study. J Sports Sci. 33(11):1149–1158. doi: 10.1080/02640414.2014.986503.

Faber IR, Oosterveld FGJ, Nijhuis-Van der Sanden MWG, Lucia A. 2014. Does an eye-hand coordination test have added value as part of talent identification in table tennis? A validity and reproducibility study. PLOS ONE. 9(1):e85657. doi: 10.1371/journal.pone.0085657.

Faul F, Erdfelder E, Buchner A, Lang A-G. 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods. 41(4):1149–1160. doi: 10.3758/BRM.41.4.1149.

Fransen J, D'Hondt E, Bourgois J, Vaeyens R, Philippaerts RM, Lenoir M. 2014. Motor competence assessment in children: convergent and discriminant validity between the BOT-2 short form and KTK testing batteries. Res Dev Disabilities. 35(6):1375–1383. doi: 10.1016/j.ridd.2014.03.011.

Gagné F. 2021. Differentiating giftedness from talent: the DMGT perspective on talent development. 2nd ed. London: Routledge. 10.4324/9781003088790.

Garn AC, Webster EK. 2021. Bifactor structure and model reliability of the test of gross motor development 3rd edition. J Sci Med Sport. 24(1):67–73. doi: 10.1016/j.jsams.2020.08.009.

Gil SM, Zabala-Lili J, Bidaurrazaga-Letona I, Aduna B, Lekue JA, Santos-Concejero J, Granados C. 2014. Talent identification and selection process of outfield players and goalkeepers in a professional soccer club. J Sports Sci. 32(20):1931–1939. doi: 10.1080/02640414.2014.964290.

Haibach P, Reid G, Collier D. 2018. Motor learning and development Vol. 2. Champaign, IL: Human Kinetics.

Hands B, McIntyre F, Parker H. 2018. The General motor ability hypothesis: an old idea revisited. Percept Mot Skills. 125(2):213–233. doi: 10.1177/0031512517751750.

Hecksteden A, Kellner R, Donath L. 2022. Dealing with small samples in football research. Sci Med Football. 6(3):389–397. doi: 10.1080/24733938.2021.1978106.

Herrmann C, Seelig H. 2017. Basic motor competencies of fifth graders. Ger J Exercise Sport Res. 47(2):110–121. doi: 10.1007/s12662-016-0430-3.

Höner O, Bergmann F, Leyhr D. in revision. "Don't forget the number 1 in soccer talent research!" – predictive validity of objectively tested motor attributes and subjectively rated position-specific skills in elite youth goalkeepers.

Höner O, Murr D, Larkin P, Schreiner R, Leyhr D. 2021. Nationwide subjective and objective assessments of potential talent predictors in elite youth soccer: an investigation of prognostic validity in a prospective study. Front Sports Act Living. 3:638227. doi: 10.3389/fspor.2021.638227.

Höner O, Votteler A, Schmid M, Schultz F, Roth K. 2015. Psychometric properties of the motor diagnostics in the German football talent identification and development programme. J Sports Sci. 33(2):145–159. doi: 10.1080/02640414.2014.928416.

Hu LT, Bentler PM. 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equation Modeling A Multidiscip J. 6(1):1–55. doi: 10.1080/10705519909540118.

Kelly AL, Eveleigh C, Bergmann F, Höner O, Braybrook K, Vahia D, Finnegan L, Finn S, Verbeek J, Jonker L, et al. 2024. International perspectives: evaluating male talent pathways from across the globe. In: Kelly AL, editor. Talent identification and development in youth soccer. Routledge; p. 228–262. 10.4324/9781032232799 .

Kiphard EJ, Schilling F. 2007. *Body coordination test for children*. Beltz: Weinheim.

Knoop M, Fernandez-Fernandez J, Ferrauti A. 2013. Evaluation of a specific reaction and action speed test for the soccer Goalkeeper. J Strength Cond Res. 27(8):2141–2148. doi: 10.1519/JSC.0b013e31827942fa.

Lakens D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Phychol. 4. doi: 10.3389/fpsyg.2013.00863.

Lämmle L, Tittlbach S, Oberger J, Worth A, Bös K. 2010. A two-level model of motor performance ability. J Exercise Sciamp Fit. 8(1):41–49. doi: 10.1016/S1728-869X(10)60006-8.

Machado e Costa F, Braga AC. 2020. Adjusting ROC Curve for Covariates with AROC R Package. In: Gervasi O Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science (Cham: Springer) 12251 doi:10.1007/978-3-030-58808-3_15

Magill R, Anderson D. 2021. Motor learning and control. Concepts and applications. New York City: McGraw Hill.

Mandrekar JN. 2010. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 5(9):1315–1316. doi: 10.1097/JTO.0b013e3181ec173d.

Marsh HW, Hau KT, Wen Z. 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct Equation Modeling A Multidiscip J. 11(3):320–341. doi: 10.1207/s15328007sem1103_2.

Matthys SPJ, Vaeyens R, Fransen J, Deprez D, Pion J, Vandendriessche J, Vandorpe B, Lenoir M, Philippaerts R. 2013. A longitudinal study of multidimensional performance characteristics related to physical capacities in youth handball. J Sports Sci. 31(3):325–334. doi: 10.1080/02640414.2012.733819.

Mokkink LB, Boers M, van der Vleuten CMB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwe CB. 2020. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a delphi study. BMC Res Methodol. 20(1):293. doi: 10.1186/s12874-020-01179-5.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. 2010a. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international delphi study. Qual Life Res. 19(4):539–549. doi: 10.1007/s11136-010-9606-8.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. 2010b. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 63(7):737–745. doi: 10.1016/j.jclinepi.2010.02.006.

Muthén LK, Muthén BO. 2017. Mplus: statistical analysis with latent variables: User's Guide (version 8). Los Angeles, CA: Authors.

O'Brien-Smith J, Tribolet R, Smith MR, Bennett KJM, Fransen J, Pion J, Lenoir M. 2019. The use of the Körperkoordinationstest für kinder in the talent pathway in youth athletes: a systematic review. J Sci Med Sport. 22(9):1021–1029. doi: 10.1016/j.jsams.2019.05.014.

Otte F, Dittmer T, West J. 2023. Goalkeeping in modern football: current positional demands and research insights. Int Sport Coaching J. 10(1):112–120. doi: 10.1123/iscj.2022-0012.

Pion JA, Fransen J, Deprez DN, Segers VI, Vaeyens R, Philippaerts RM, Lenoir M. 2015. Stature and jumping height are required in female volleyball, but motor coordination is a key factor for future elite

success. J Strength Cond Res. 29(6):1480–1485. doi: 10.1519/jsc.0000000000000778.

Pion J, Segers VI, Fransen J, Debuyck G, Deprez DN, Haerens L, Vaeyens R, Philippaerts RM, Lenoir M. 2015. Generic anthropometric and performance characteristics among elite adolescent boys in nine different sports. Eur J Sport Sci. 15(5):357–366. doi: 10.1080/17461391.2014.944875.

Portney LG, Watkins MP. 2015. Foundations of clinical research: applications to practice. 3rd ed. Philadelphia: F.A. Davis Company.

Preckel F, Golle J, Grabner R, Jarvin L, Kozbelt A, Müllensiefen D, Olszewski-Kubilius P, Schneider W, Subotnik R, Vock M, et al. 2020. Talent development in achievement domains: a psychological framework for within- and cross-domain research. Perspectives On Psychological Sci a J Assoc For Psychological Sci. 15(3):691–722. doi: 10.1177/1745691619895030.

R Core Team. 2021. R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. https://www.R-project.org/.

Rebelo-Gonçalves R, Figueiredo AJ, Coelho-E-Silva MJ, Tessitore A. 2016. Assessment of technical skills in young soccer goalkeepers: reliability and validity of two goalkeeper-specific tests. J Sports Sciamp Med. 15(3):516–523.

Rommers N, Mostaert M, Goossens L, Vaeyens R, Witvrouw E, Lenoir M, D'Hondt E. 2019. Age and maturity related differences in motor coordination among male elite youth soccer players. J Sports Sci. 37(2):196–203. doi: 10.1080/02640414.2018.1488454.

Sieghartsleitner R, Zuber C, Zibung M, Charbonnet B, Conzelmann A. 2019. Talent selection in youth football: technical skills rather than general motor performance predict future player status of football talents. Curr Issues Sport Sci. 4:011. doi: 10.15203/CISS_2019.011.

Sieghartsleitner R, Zuber C, Zibung M, Conzelmann A. 2019. Science or coaches' eye?–Both! Beneficial Collaboration of Multidimensional Measurements and Coach Assessments for Efficient Talent Selection in Elite Youth Football. J Sports Sci Med. 18(1):32–43.

Tavakol M, Dennick R. 2011. Making sense of Cronbach's alpha. Int J Med Educ. 2:53–55. doi: 10.5116/ijme.4dfb.8dfd.

Thompson B. 2004. Exploratory and confirmatory factor analysis: understanding concepts and applications. American Psychological Association. 10.1037/10694-000.

Turvey MT. 1990. Coordination. Am Psychologist. 45(8):938–953. doi: 10.1037/0003-066X.45.8.938.

Utesch T, Bardid F. 2019. Motor Competence. In: Schinke DH, J. R Strauß B, editors. Dictionary of Sport Psychology. Sport, Exercise, and Performing Arts. Amsterdam: Elsevier; p. 186.

Vahia D, Kelly AL. 2024. The Goalkeeper: highlighting the position data gap in talent identification and development. In: Kelly AL, editor. Talent identification and development in youth soccer. Routledge; p. 294–316. 10.4324/9781032232799.

Vandorpe B, Vandendriessche JB, Vaeyens R, Pion J, Lefevre J, Philippaerts RM, Lenoir M. 2012. The value of a non-sport-specific motor test battery in predicting performance in young female gymnasts. J Sports Sci. 30(5):497–505. doi: 10.1080/02640414.2012.654399.

Weir JP. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength & Cond Res. 19(1):231–240. doi: 10.1519/00124278-200502000-00038.

West J. 2018. A review of the key demands for a football goalkeeper. Int J Sports Sciamp Coaching. 13(6):1215–1222. doi: 10.1177/1747954118787493.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York.

Williams A, Ford P, Drust B. 2020. Talent identification and development in soccer since the millennium. J Sports Sci. 38(11–12):1199–1210. doi: 10.1080/02640414.2020.1766647.