Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Principles for designing game-based assessments of individual performance in team sports

Robbie Wilson

r.wilson@uq.edu.au

University of Queensland - Saint Lucia Campus: The University of Queensland https://orcid.org/0000-0002-0116-5427

Andrew H. Hunter

The University of Queensland - Saint Lucia Campus: The University of Queensland

Nicholas Smith

The University of Queensland - Saint Lucia Campus: The University of Queensland

Thiago Camata

The University of Queensland - Saint Lucia Campus: The University of Queensland

Mathew S. Crowther

The University of Sydney

Research Article

Keywords: Game-based assessment, talent identification, reliability, individual performance, small-sided games, football, replication, player assessment

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11172785/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background

Game-based assessments capture how individual capabilities are expressed against opponents and teammates, but these interactions also introduce variation that can obscure consistent differences among athletes. We used repeated one-versus-one (1v1) contests in youth football to determine how much replication is required to estimate individual competitive performance reliably and whether this depends on the population assessed.

Methods

Across five player groups, we quantified repeatability of attacking, defensive and overall 1v1 performance and examined how reliability changed with replication. We also tested differences between adjacent age cohorts and used empirically parameterised simulations to determine how replication and between-player heterogeneity affected recovery of expected player performance.

Results

Single-contest repeatability was low for attacking (R = 0.12–0.18), defensive (R = 0.08–0.21) and overall performance (R = 0.16–0.26), but increased substantially with aggregation. Overall performance reached an effective repeatability of 0.80 after approximately 12–22 contests. Older players showed higher attacking and defensive performance across three adjacent age-cohort comparisons. In simulations, repeated contests progressively recovered expected performance across opponents. Under reference heterogeneity, at least 90% of simulated populations achieved $r \geq 0.80$ after 15 contests at $\rho = 0.3$. The same criterion required nine contests when between-player variance was doubled but 27 when it was halved.

Conclusions

There is no universal number of contests required for reliable game-based assessment. Replication requirements depend on differences among the athletes being assessed, so assessment intensity should be determined from the task, population and precision required.

Key Points

• Individual competitive outcomes can provide reliable estimates of player performance, but single contests are noisy and repeated sampling across competitive interactions is required.

• The amount of replication required depends strongly on between-player heterogeneity: more similar athletes require substantially more observations to distinguish reliably.

• Game-based assessment should therefore be designed around the task, population and required precision rather than using a fixed number of games or contests across assessment contexts.

Introduction

Quantitative assessment of individual athletes is fundamental to talent identification and player development across team sports, supporting selection, monitoring and feedback on individual strengths and weaknesses [1–3]. Most assessments measure physical, technical, perceptual-cognitive or psychological capabilities individually under standardised conditions, allowing external variation to be controlled and differences to be attributed more directly to the individual athlete [1–3]. Yet performance in team sports ultimately depends on how these capabilities are expressed through interactions with teammates and opponents under changing spatial, temporal and tactical constraints [4]. At the other end of this assessment continuum is match analysis, which provides highly representative information about athletes within competition but with far less control over the context in which performance occurs [5]. Because opportunities and outcomes in matches depend on tactical roles, game states, teammates and opponents, individual performance can vary substantially among matches [5–8]. Individual assessment in team sports therefore presents a fundamental trade-off: greater experimental control can improve the isolation and repeatability of individual metrics, whereas greater competitive representativeness makes it increasingly difficult to separate the athlete's contribution from the context in which it occurs.

Game-based assessments under more controlled conditions provide an intermediate approach. However, individual performance within games can be quantified in fundamentally different ways. Established game-performance instruments commonly evaluate observer-defined components such as decision making, skill execution, support and tactical behaviour [9, 10], while the broader game-based coaching literature has similarly focused heavily on technical execution, decision making and tactical awareness [11, 12]. These measures can provide important information about *how* an athlete performs, but they are conceptually different from measuring whether the athlete actually succeeds in the competitive interaction. For talent identification, objective competitive outcomes have an important attraction: the performance event itself can often be defined unambiguously. An attacker either successfully beats a defender or does not; a defender prevents progression or does not; a scoring opportunity succeeds or fails. Comparatively few studies have used such competitive outcomes directly to quantify individual performance. Fenner et al. [13], for example, accumulated points from match results and goals scored across repeated 4v4 football games, while Wilson et al. [14] quantified individual performance across repeated 3v3 games from goals scored, goals conceded and net game outcomes while systematically varying teammates and opponents. However, even where objective competitive outcomes have been used, little attention has been given to determining how much repeated competitive sampling is required before accumulated success provides a stable estimate of the individual athlete.

An objectively measured competitive outcome does not, however, provide an unambiguous estimate of the individual who produced it. Success in team sports is inherently relational: an individual's observed outcome depends partly on the other athletes involved. An attacker may succeed against one defender but fail against another, while in multi-player games individual outcomes additionally depend on the teammates alongside whom the athlete performs. This attribution problem is well recognised in team-sport analytics, where methods such as adjusted plus-minus attempt to estimate individual contributions to objectively observed scoring outcomes while accounting for the teammates and opponents present, yet individual estimates can remain influenced by the particular combinations of players sampled [15]. Even when the same athletes compete again, a limited number of opportunities can produce different outcomes from one contest to the next. The measurement problem is therefore not necessarily uncertainty about what happened in the game, but uncertainty about how much that observed outcome tells us about the individual. Competitive situations can nevertheless be structured so that these contextual sources of variation are deliberately sampled. Players can be observed repeatedly against different opponents and, in more complex games, alongside different teammates, allowing performance consistently associated with the focal athlete to be progressively distinguished from performance arising from particular matchups or combinations. This follows the broader principle of performance assessment that the generalisability of an individual's score depends on sampling across the conditions under which performance is observed [16, 17]. The resulting estimates remain relative to the competitive population in which they were obtained rather than directly comparable across different groups or populations.

This attribution problem makes replication fundamental to outcome-based competitive assessment. A player's poor performance in one contest might reflect lower underlying ability, an unusually strong opponent or simply an unsuccessful set of competitive opportunities on that occasion. In multi-player games, the problem expands further because outcomes can also depend on teammate quality and particular combinations of players. A single observation therefore provides limited information about which component belongs consistently to the focal athlete. More generally, studies that have repeated identical or closely matched small-sided games show that the reliability of game-derived measures can vary substantially among performance variables and game formats [18, 19]. Evidence from football using individual competitive outcomes illustrates more directly how repeated sampling can resolve this uncertainty. In *rondos*—possession games in which a numerical majority attempts to retain the ball against one or more defenders—passing success shows low repeatability from a single bout in 3v1 formats but substantially greater repeatability when multiple bouts are aggregated [20]. Likewise, objective behavioural and scoring outcomes from repeated 3v3 games can vary considerably as teammate and opponent combinations change, yet stable differences among players become increasingly apparent when performance is aggregated across games [14]. Game-based competitive outcomes are therefore not necessarily poor individual metrics because single observations are noisy; rather, reliable information about individuals may emerge by repeatedly sampling the competitive contexts in which success is expressed.

Football provides a particularly useful model system for examining this measurement problem because assessment can be simplified progressively from competitive matches to small-sided games, possession tasks and direct player-versus-player contests [2, 21–23]. Direct 1v1 contests provide the simplest case for studying the attribution of objective competitive outcomes to individuals because each outcome is generated by only one attacker and one defender, and comparable attacker–defender interactions form identifiable competitive sub-phases in other invasion sports such as basketball [24, 25]. Attackers must integrate technical skill, movement, deception and decision making against an actively responding opponent, while defenders must anticipate these actions, control space and prevent progression or regain possession [24–27]. Success or failure can be recorded directly, opponents can be systematically varied across repeated contests, and analytical approaches can explicitly account for differences in opponent quality. Previous work has demonstrated substantial individual variation in football 1v1 performance and associations between football-specific capabilities, particularly dribbling and sprinting, and competitive success [26, 27]. The critical measurement question is therefore how many contests are required before consistent differences among players can be distinguished from variation among individual competitive outcomes?

There is unlikely to be a single answer to this question. The number of observations required to distinguish athletes should depend on the magnitude of stable between-athlete differences relative to variation occurring from contest to contest. This follows a general principle of reliability: the ability to distinguish consistently among individuals depends on the magnitude of stable between-individual differences relative to other sources of variation, while repeated observations increase the precision with which those differences can be estimated [16, 28–30]. Importantly, measures of relative reliability are themselves sensitive to the heterogeneity of the population being assessed, so the same measurement procedure can provide different reliability when applied to populations containing broader or narrower ranges of performance [28, 29]. When athletes differ substantially in ability, those differences should provide a strong signal and relatively few observations may be sufficient. When athletes are more similar, the same contest-to-contest variation may obscure smaller underlying differences, requiring more observations before individuals can be reliably distinguished. The replication required to estimate individual competitive performance should therefore increase as the athletes being compared become more similar in ability.

This prediction has a direct implication for talent-identification pathways. Talent identification commonly operates as a staged process in which larger initial candidate populations are progressively reduced through successive screening and selection decisions [1, 22, 31]. If these selection stages successfully discriminate among athletes on characteristics relevant to performance, they should do more than reduce the number of candidates, they should also progressively reduce the magnitude of the differences among those who remain. Early screening may therefore involve populations spanning a relatively broad range of current abilities, whereas later decisions may involve much smaller groups of more closely matched athletes. Consequently, later-stage decisions may require more observations per athlete, not fewer, despite involving fewer athletes, because previous selection has reduced the between-athlete variation against which subsequent differences must be detected. Assessment intensity should

therefore increase as selection narrows the differences among the remaining candidates, even as the total number of athletes being assessed decreases.

Here, we used repeated 1v1 contests in youth football as an empirical model system to determine how replication affects the reliability and discrimination of objective individual competitive-performance estimates. Across five independent player groups, we first quantified between-player variation in attacking, defensive and overall competitive success and determined how reliability changed as contests were aggregated. We then tested whether opponent-adjusted 1v1 performance discriminated between adjacent age cohorts and examined how consistently the same players performed when they encountered one another again. These empirical analyses were complemented by generative simulations in which each player's expected attacking, defensive and overall performance was specified by the model, allowing us to determine how accurately performance estimated from a limited number of contests recovered these underlying differences as replication increased. Finally, we systematically manipulated between-player heterogeneity while holding the assessment structure constant and tested the robustness of the results to uncertainty in the association between attacking and defensive abilities. We predicted that repeated observations would progressively improve recovery of individual competitive performance, but that substantially greater replication would be required as differences among players narrowed. Thus, rather than seeking a universal number of contests required for reliable assessment, we test the broader prediction that the sampling intensity required to estimate competitive performance depends on the heterogeneity of the athletes being compared. Although parameterised using football 1v1 contests, this approach provides a framework for converting objectively measured but context-dependent competitive outcomes into reliable estimates of individual performance.

Methods

Study design

We combined empirical analyses of repeated one-versus-one (1v1) football contests with empirically parameterised simulations to determine how replication influences the reliability and accuracy of individual competitive-performance estimates. Five player groups were analysed independently because performance was defined relative to the opponents available within each assessment population. Groups 1–3 comprised academy players within the same academy tested on separate assessment days, while Groups 4 and 5 comprised female and male youth players from the same academy.

The final empirical dataset comprised Group 1 (53 players, 345 contests), Group 2 (57 players, 351 contests), Group 3 (61 players, 435 contests), Group 4 (29 players, 127 contests) and Group 5 (22 players, 136 contests), giving 222 player-by-group records and 1,394 contests in total. Player-level unadjusted summaries included attack successes per contest, defensive successes per contest, net successes per contest, the standard deviation and median of successes, and mean absolute contest margin. Written informed consent was obtained from players and their parent or legal guardian. Ethical

approval was granted by The University of Queensland Human Research Ethics Committee (2021/HE001974).

One-versus-one assessment

Competitive performance was assessed using repeated game-representative 1v1 contests based on a previously described protocol [27]. Contests were conducted on a 25 × 13 m playing area with a 2 × 13 m scoring zone at one end. Each contest involved two players who alternated between attacking and defensive roles. An attacking opportunity commenced with the defender passing the ball to the attacker from the edge of the scoring zone, after which the attacker attempted to enter the scoring zone while maintaining possession. Attacking success was recorded when the attacker contacted the ball within the scoring zone. Defensive success was recorded when the defender dispossessed the attacker, forced the ball out of play, or prevented the attacker from entering the scoring zone within the 15-s time limit.

Players alternated roles until each had completed three attacking and three defensive opportunities against the same opponent. For each player within each contest, attacking performance was quantified as the number of successful attacks across three opportunities (0–3), defensive performance as the number of successful defensive outcomes across three opportunities (0–3), and overall performance as net goal difference, calculated as the focal player's attacking successes minus those of the opponent (−3 to +3).

Opponent-adjusted player performance

Opponent-adjusted performance was estimated separately within each empirical group using crossed binomial generalised linear mixed models with a logit link. Each contest contributed two directed observations, one for each player in the attacker role. Player identity was represented through crossed attacker and defender random intercepts, allowing the focal player's scoring propensity to be separated from the strength of the opponent encountered. The attacker random intercept represented latent attacking ability. The defender random intercept represented scoring allowed and was sign-reversed so that larger values represented better defensive ability. Model-based opponent-adjusted performance was then expressed on the observable contest scale. For each player, expected attacking successes were predicted against the relevant opponent population from the fitted crossed attacker–defender model. Expected goals conceded were calculated analogously from the scoring probabilities of the opponents when attacking the focal player. Opponent-adjusted overall performance was defined as the player's mean expected goals scored minus mean expected goals conceded across the relevant opponent population. Thus, overall performance represented the expected competitive net outcome against the population of available opponents rather than the arithmetic sum of the two latent random effects. Players were ranked using both raw net performance and this opponent-adjusted expected net performance to quantify the extent to which accounting for opponent strength altered estimates of relative player performance. Model convergence and singular-fit diagnostics were retained for all fitted models.

Reliability of repeated performance estimates

Reliability was assessed using complementary approaches that quantified the stability of performance at different levels of replication. First, single-contest repeatability was estimated separately for attacking, defensive and overall performance within each empirical group. The expected effective repeatability after aggregating k contests was then calculated from the single-contest repeatability, R, using the Spearman–Brown relationship: $R_{eff} = kR/[1 + (k - 1)R]$. This provided the expected reliability of an assessment comprising multiple contests and allowed the number of contests required to attain effective repeatabilities of 0.70, 0.80 and 0.90 to be determined. Uncertainty in effective repeatability was propagated from 5,000-resample bootstrap confidence intervals for single-contest repeatability.

Second, the stability of opponent-adjusted player performance was evaluated empirically using 5,000 random split-half partitions of contests within each group. For each partition, whole contests were randomly divided as evenly as possible between two subsets, and the crossed attacker–defender model was independently refitted to each half. Opponent-standardised expected net performance was then recalculated separately within each half from the corresponding fitted model. Agreement between halves was assessed for players represented in both subsets using Pearson correlations for attacking, defensive and overall performance and Spearman rank correlations for overall performance. These correlations quantified agreement between two half-sample assessments. Pearson split-half correlations were subsequently converted using the Spearman–Brown correction, $R_{SB} = 2r/(1+r)$, to estimate the reliability expected for the complete assessment containing both halves. Median estimates across successful partitions were used as the primary measures of reliability, with the 2.5th and 97.5th percentiles providing resampling intervals. Singular half-sample fits were recorded.

The aggregation and split-half analyses therefore addressed related but distinct questions: the first estimated how expected reliability changes with the number of repeated contests, whereas the second assessed the empirical stability of opponent-adjusted player ratings across independent subsets of the observed contests.

Age-category comparisons

Prespecified adjacent-cohort comparisons were U10 versus U11 in Group 1, U11 versus U12 in Group 3 and U12 versus U13 in Group 2. Players outside each prespecified comparison were excluded from the corresponding age analysis. Age information was successfully linked for 53/53, 56/57 and 57/61 players in Groups 1–3, respectively; the prespecified adjacent-cohort comparisons included 53, 56 and 52 players. Formal inference was conducted at the contest level using crossed binomial GLMMs containing age category as a fixed effect, with the younger cohort as the reference, while retaining the crossed player and opponent structure. Attacking and defensive age effects were expressed as odds ratios (ORs) for the older relative to the younger cohort, with Wald 95% confidence intervals and two-sided *P* values. For descriptive interpretation, opponent-adjusted attacking performance, defensive performance and overall expected net performance were additionally summarised using age-cohort means, mean differences, Hedges' *g* and area under the receiver-operating-characteristic curve (AUC). Overall expected

net performance was calculated as the model-predicted expected net outcome against the relevant opponent population. These player-level summaries were descriptive and were not used as the basis of formal statistical inference.

Consistency of repeated encounters between the same players

Rematch consistency was assessed using unordered player pairs that completed at least two valid contests (random occurrences) within the same assessment group. Contests were ordered according to their occurrence in the source data, and the primary analysis compared the first and second encounters. Third encounters were retained in the data inventory but were excluded from the primary first-versus-second analysis.

Directed attacking scores were oriented so that the same player was treated as the attacker in both encounters, while overall performance was represented by consistently oriented net margin. First-versus-second agreement was summarised using Pearson and Spearman correlations, exact agreement and mean absolute differences. Uncertainty was estimated using 5,000 non-parametric cluster-bootstrap resamples of complete unordered pairs, with pooled resampling stratified by empirical group.

Model-based rematch consistency was additionally evaluated using binomial mixed models containing an indicator for the second encounter, crossed attacker and defender random intercepts, and a directed attacker–defender matchup random intercept. Matchup models were compared with reduced attacker–defender models using log likelihood, AIC, singularity diagnostics and a boundary-adjusted likelihood-ratio test.

Simulation of repeated 1v1 assessments

We used an empirically parameterised generative simulation to determine how increasing replication affected recovery of underlying player performance and whether the required replication depended on between-player heterogeneity. The simulations were parameterised from the crossed attacker–defender binomial models fitted to the five empirical groups. For the primary simulation, median values across the empirical groups were used as the reference parameterisation: the logit-scale intercept was $\mu = -0.461$, corresponding to a baseline scoring probability of 0.387 per opportunity, while the between-player variances were 0.220 for attacking ability and 0.278 for defensive ability.

Each simulated population contained 50 players. For player i, latent attacking ability A_i and defensive ability D_i were generated from a bivariate normal distribution with means of zero, variances σ_A^2 and σ_D^2, and covariance $\rho\sigma_A\sigma_D$. Because the empirical crossed random-effects models estimated separate attacker and defender variances but did not directly estimate their covariance, the attack–defence correlation ρ was treated as a sensitivity parameter, with simulations conducted at $\rho = 0$, 0.3 and 0.6.

For a directed contest in which player i attacked player j, scoring probability for each opportunity was generated according to $\text{logit}(p_{ij}) = \mu + A_i - D_j$. Each player received three scoring opportunities

against the opponent, matching the empirical assessment, and goals scored were generated from a Binomial(3, p_ij) distribution.

Players were randomly paired within each simulated round so that every player completed one contest per round. Forty rounds were generated for each population. Performance was evaluated cumulatively from 3 to 40 contests per player using the same simulated population and contest sequence, rather than generating an independent population at each replication level. Thus, performance after 20 contests contained the first 10 contests used to estimate performance at 10 contests plus 10 additional contests.

Latent attacking and defensive abilities represented the model-specified underlying attacking and defensive performance. Model-specified overall performance was defined on the observable contest scale rather than simply as the sum of the two latent effects. For each player, expected goals scored and expected goals conceded were calculated against every other player in the simulated population using the generating model, and underlying overall performance was defined as mean expected goals scored minus mean expected goals conceded.

At each replication level k, observed attacking performance was calculated as cumulative goals scored divided by k, defensive performance as cumulative unsuccessful opponent scoring opportunities divided by k, and overall performance as cumulative net goals per contest, (goals scored − goals conceded)/k. Recovery was quantified as the Pearson correlation, r, between observed and model-specified player performance across the 50 players in each simulated population.

Effects of between-player heterogeneity

The primary simulation experiment tested whether replication requirements depended on the magnitude of between-player differences. Both reference variance components were multiplied by 0.5, 1.0 or 2.0 to represent low, reference and high between-player heterogeneity, respectively. Each heterogeneity treatment was crossed with the three assumed attack–defence correlations (ρ = 0, 0.3 and 0.6), producing nine simulation conditions.

For each condition, 5,000 independent simulated populations were generated. At each level of replication, we calculated the mean and median recovery correlation, its 5th and 95th percentiles, and the proportion of simulations attaining recovery correlations of at least 0.70, 0.80 and 0.90. The primary replication criterion was the minimum number of contests at which at least 90% of simulated populations achieved r ≥ 0.80. This criterion incorporated both the magnitude and consistency of performance recovery rather than relying on mean recovery alone.

As a complementary sensitivity analysis, simulations were repeated using the empirical intercept and attacking and defensive variance estimates from each of the five empirical groups separately. These simulations retained the 50-player, 3–40-contest design and were repeated at ρ = 0, 0.3 and 0.6, allowing replication requirements to be compared across the range of heterogeneity observed empirically.

Software and reproducibility

Analyses were conducted in R 4.6.1. Empirical mixed-model, bootstrap, rematch and age analyses used lme4 and associated data-processing and model-utility packages, with figures produced using ggplot2. The simulation workflow used MASS 7.3-66 for multivariate-normal generation, dplyr 1.2.1, tidyr 1.3.2, purrr 1.2.2 and readr 2.2.0 for data manipulation and output, and ggplot2 4.0.3 for graphical presentation. Random-number seeds were fixed for stochastic analyses, and input-file checksums, contest audits, model diagnostics, bootstrap outputs and simulation draws were retained for reproducibility.

Results

Sample composition and opponent adjustment

Across the five empirical groups, 222 player-by-group records contributed 1,394 contests (Table 1). Draws accounted for 22.5–30.6% of contests. All five crossed attacker–defender models fitted to the complete datasets converged without singular fits. Raw and opponent-adjusted net performance were strongly associated (Pearson r = 0.963–0.991; Spearman ρ = 0.971–0.982), indicating that opponent adjustment retained the broad ordering of players while modifying individual rankings (Table 1). Mean absolute rank changes ranged from 0.95 positions in Group 5 to 3.31 positions in Group 3, with a maximum individual rank change of 13 positions in Group 3.

Table 1
Sample characteristics and effects of adjusting individual 1v1 performance for opponent strength.

Group	Players	Contests	Drawn (%)	SD expected net	Pearson r	Spearman ρ	Mean |rank Δ|	Max rank Δ
1	53	345	24.1	0.602	0.987	0.976	2.60	8
2	57	351	22.5	0.617	0.980	0.982	2.54	8
3	61	435	30.6	0.391	0.963	0.971	3.31	13
4	29	127	25.2	0.352	0.972	0.973	1.45	5
5	22	136	25.0	0.434	0.991	0.975	0.95	5

Note. Expected net = opponent-standardised expected net performance, calculated as mean expected goals scored minus mean expected goals conceded across the relevant opponent population. Pearson *r* and Spearman ρ quantify the association between raw net performance and opponent-adjusted expected net performance. Rank Δ represents the change in player rank after accounting for opponent strength.

Reliability of repeated performance estimates

Single-contest repeatability was low across the five groups, ranging from R = 0.12–0.18 for attacking performance, R = 0.08–0.21 for defensive performance and R = 0.16–0.26 for overall performance. Consequently, individual contests provided relatively imprecise estimates of stable between-player

differences. However, expected reliability increased strongly as contests were aggregated (Fig. 1). Based on the single-contest estimates and Spearman–Brown relationship, overall performance reached an effective repeatability of 0.80 after approximately 12–22 contests, depending on the empirical group.

The independent split-half analysis provided complementary evidence that opponent-adjusted performance estimates were stable across independent subsets of contests (Table 2). Median Spearman–Brown-corrected full-sample reliability for opponent-standardised expected net performance was 0.81 in Groups 1 and 2, 0.74 in Group 3, 0.70 in Group 4 and 0.73 in Group 5. Attacking reliability ranged from 0.60 to 0.74 and defensive reliability from 0.59 to 0.77. All 5,000 partitions produced successful overall correlations in Groups 1–3; 4,977 were successful in Group 4 and 4,999 in Group 5. Singular half-sample fits were absent from Groups 1–2, occurred in 25 of 5,000 partitions in Group 3, and were more frequent in the smaller Group 4 (1,257/5,000) and Group 5 (247/5,000) datasets.

Table 2
Full-sample reliability of opponent-adjusted 1v1 performance estimated from independent split-half assessments.

Group	Measure	Median full-sample reliability	95% resampling interval
1	Attack	0.739	0.617–0.829
	Defence	0.724	0.588–0.821
	Overall	0.814	0.731–0.877
2	Attack	0.680	0.532–0.786
	Defence	0.770	0.669–0.846
	Overall	0.805	0.720–0.870
3	Attack	0.663	0.521–0.775
	Defence	0.591	0.417–0.731
	Overall	0.736	0.623–0.824
4	Attack	0.598	0.322–0.787
	Defence	0.621	0.360–0.797
	Overall	0.698	0.480–0.841
5	Attack	0.665	0.307–0.830
	Defence	0.712	0.493–0.855
	Overall	0.728	0.467–0.867

Overall = opponent-standardised expected net performance. Values are median Spearman–Brown-corrected estimates of full-sample reliability across 5,000 random split-half partitions; 95% intervals are the 2.5th–97.5th percentiles across successful partitions.

Differences between adjacent age cohorts

Older cohorts showed higher attacking and defensive performance in all three prespecified comparisons (Table 3). The U11 players had 1.69 times the odds of attacking success relative to U10 players (95% CI 1.17–2.43, $P = 0.0051$) and 2.17 times the odds of defensive success (95% CI 1.62–2.91, $P < 0.0001$). The U12 players exceeded U11 players in attack (OR = 1.73, 95% CI 1.31–2.28, $P < 0.0001$) and defence (OR = 1.49, 95% CI 1.13–1.95, $P = 0.0041$), and the U13 players also exceeded U12 players in attack (OR = 1.83, 95% CI 1.35–2.48, $P = 0.0001$) and defence (OR = 2.66, 95% CI 1.92–3.70, $P < 0.0001$).

Table 3
Differences in attacking and defensive 1v1 performance between adjacent age cohorts.

Comparison	Players	Contests	Attack OR	95% CI attack	P attack	Defence OR	95% CI defence	P defence
U10 vs U11	53	345	1.69	1.17–2.43	0.0051	2.17	1.62–2.91	< 0.0001
U12 vs U13	56	337	1.83	1.35–2.48	0.0001	2.66	1.92–3.70	< 0.0001
U11 vs U12	52	335	1.73	1.31–2.28	< 0.0001	1.49	1.13–1.95	0.0041

Note. Odds ratios (ORs) represent performance of the older cohort relative to the younger cohort. Formal inference was based on contest-level crossed binomial GLMMs accounting for player and opponent identity.

Descriptive opponent-adjusted overall expected net performance was also higher in the older cohort in each comparison: the older-minus-younger difference was 0.611 expected net goals in the u10/u11 comparison (Hedges' $g = 1.15$; AUC = 0.81), the 0.350 in u11/u12 comparison, and 0.763 in the u12/u13 comparison ($g = 1.57$; AUC = 0.85), ($g = 0.97$; AUC = 0.75) (Table 4; Fig. 2). Thus, the contest-level models consistently discriminated adjacent annual age cohorts in attacking and defensive performance, while the opponent-standardised overall expected net performance showed the same directional developmental difference.

Table 4

Descriptive differences in opponent-adjusted 1v1 performance between adjacent age cohorts. Group 1: U10 (N = 26) vs U11 (n = 27); Group 3: U11 (N = 22) vs U12 (n = 30); Group 2: U12 (N = 30) vs U13 (n = 26);

Comparison	Measure	Mean young	Mean old	Difference	Hedges g	AUC
U10 vs U11	Attack	-0.190	0.203	0.393	0.77	0.691
	Defence	-0.286	0.259	0.545	1.41	0.869
	Overall	-0.476	0.462	0.938	1.15	0.815
U11 vs U12	Attack	-0.176	0.130	0.306	0.98	0.736
	Defence	-0.106	0.099	0.205	0.77	0.714
	Overall	-0.282	0.230	0.511	0.98	0.752
U12 vs U13	Attack	-0.171	0.228	0.399	1.05	0.771
	Defence	-0.350	0.412	0.761	1.60	0.867
	Overall	-0.521	0.639	1.160	1.57	0.855

Note. Attack and Defence represent opponent-adjusted player performance from the crossed attacker–defender models. Overall represents opponent-standardised expected net performance. Differences are calculated as older minus younger cohort. Hedges' *g* and AUC are descriptive player-level summaries; formal age-group inference is reported in Table 3.

Consistency of repeated encounters

Across the five empirical groups, 222 player-by-group records contributed 1,394 contests. Draws accounted for 22.5–30.6% of contests. All five crossed attacker–defender models fitted to the complete datasets converged without singular fits. Raw net performance and opponent-adjusted expected net performance were strongly associated (Pearson r = 0.963–0.991; Spearman ρ = 0.971–0.982), indicating that accounting for opponent strength retained the broad ordering of players while modifying individual rankings. Mean absolute rank changes ranged from 0.95 positions in Group 5 to 3.31 positions in Group 3, with a maximum individual rank change of 13 positions in Group 3.

Recovery of underlying performance in simulations

Recovery of model-specified player performance increased strongly with repeated assessment. Under reference empirical heterogeneity and the intermediate attack–defence correlation (ρ = 0.3), mean recovery of overall performance increased from r = 0.586 after three contests to 0.798 after 10 contests, 0.852 after 15 contests, 0.882 after 20 contests and 0.936 after 40 contests. The distribution of recovery correlations also narrowed with increasing replication: the 5th–95th percentile interval decreased from 0.426–0.723 after three contests to 0.784–0.907 after 15 contests and 0.906–0.960 after 40 contests.

The probability of attaining r ≥ 0.80 increased from 0.2% after three contests to 52.7% after 10 contests, 80.2% after 13 contests and 90.7% after 15 contests. Consequently, under the reference parameterisation and ρ = 0.3, 15 contests per player were required for at least 90% of simulated populations to recover underlying overall performance at r ≥ 0.80.

Effects of between-player heterogeneity

The number of contests required for accurate recovery depended strongly on between-player heterogeneity (Table 5). At ρ = 0.3, the criterion of at least 90% of simulated populations attaining r ≥ 0.80 was reached after 9 contests under high heterogeneity, 15 contests under reference heterogeneity and 27 contests under low heterogeneity. The corresponding probabilities at these thresholds were 0.905, 0.907 and 0.905, respectively.

Table 5
Replication required to recover model-specified overall performance across levels of between-player heterogeneity.

Heterogeneity	Variance multiplier	ρ	Contests: r ≥ 0.70	Contests: r ≥ 0.80	Contests: r ≥ 0.90
Low	0.5	0.0	20	34	Not reached
Low	0.5	0.3	16	27	Not reached
Low	0.5	0.6	13	22	Not reached
Reference	1.0	0.0	11	19	Not reached
Reference	1.0	0.3	9	15	35.0
Reference	1.0	0.6	8	13	29.0
High	2.0	0.0	7	11	26.0
High	2.0	0.3	6	9	21.0
High	2.0	0.6	5	8	18.0

Note. Values are the minimum number of contests at which ≥ 90% of 5,000 simulated populations achieved the specified correlation between observed and model-specified overall performance. Low, reference and high heterogeneity correspond to variance multipliers of 0.5, 1.0 and 2.0, respectively. “Not reached” indicates that the criterion was not achieved within 40 contests.

Under high heterogeneity, mean recovery was already r = 0.834 after eight contests and increased to 0.849 after nine. In contrast, under low heterogeneity, mean recovery was only 0.706 after 10 contests and 0.815 after 20 contests; the 90% probability criterion was not reached until 27 contests, when mean recovery was r = 0.853. Thus, substantially more repeated observations were required when differences among players were small relative to contest-to-contest variation.

Sensitivity to attack–defence correlation

The assumed correlation between attacking and defensive abilities also influenced performance recovery, although its effect was smaller than that of between-player heterogeneity (Table 6). Under reference heterogeneity, the number of contests required for at least 90% of simulated populations to attain r ≥ 0.80 declined from 19 contests at ρ = 0 to 15 at ρ = 0.3 and 13 at ρ = 0.6.

Table 6
Replication required to recover model-specified overall performance in simulations parameterised from each empirical group.

Empirical group	ρ	Contests: r ≥ 0.80
1	0.0	14
	0.3	12
	0.6	10
2	0.0	13
	0.3	11
	0.6	9
3	0.0	26
	0.3	21
	0.6	17
4	0.0	19
	0.3	15
	0.6	13
5	0.0	25
	0.3	19
	0.6	16

Note. Values are the minimum number of contests at which ≥ 90% of 5,000 simulated populations achieved *r* ≥ 0.80 between observed and model-specified overall performance. Simulations were conducted separately at attack–defence correlations (ρ) of 0, 0.3 and 0.6.

A more stringent recovery criterion of r ≥ 0.90 required considerably greater replication. Under reference heterogeneity, at least 90% of simulations achieved this criterion after 35 contests at ρ = 0.3 and 29 contests at ρ = 0.6, whereas the criterion was not reached within 40 contests when ρ = 0. Under high heterogeneity, the corresponding requirements were 26, 21 and 18 contests for ρ = 0, 0.3 and 0.6, respectively. Under low heterogeneity, the 90% probability criterion for r ≥ 0.90 was not attained within 40 contests at any value of ρ.

Replication requirements across the empirical populations

Simulations parameterised separately from each empirical group confirmed that replication requirements varied among the observed player populations (Table 7). At the intermediate attack–defence correlation of ρ = 0.3, at least 90% of simulations attained r ≥ 0.80 after 11 contests for Group 2, 12 for Group 1, 15 for Group 5, 19 for Group 4 and 21 for Group 3. Across ρ = 0, 0.3 and 0.6, respectively, the corresponding requirements were 14, 12 and 10 contests for Group 1; 13, 11 and 9 for Group 2; 26, 21 and 17 for Group 3; 19, 15 and 13 for Group 4; and 25, 19 and 16 for Group 5.

Table 7
Recovery of model-specified overall performance with increasing replication under reference heterogeneity (ρ = 0.3).

Contests	Mean r	5th percentile	95th percentile	Pr(r ≥ 0.80)
3	0.586	0.426	0.723	0.002
5	0.684	0.555	0.792	0.038
10	0.798	0.708	0.871	0.527
13	0.834	0.759	0.895	0.802
15	0.852	0.784	0.907	0.907
20	0.882	0.829	0.926	0.989
30	0.917	0.878	0.948	1.000
40	0.936	0.906	0.960	1.000

Note. Values summarise 5,000 simulated populations. Percentiles represent the 5th and 95th percentiles of the distribution of recovery correlations. Pr(*r* ≥ 0.80) is the proportion of simulated populations in which observed overall performance correlated with model-specified overall performance at *r* ≥ 0.80.

Taken together, the empirical and simulation analyses produced closely corresponding estimates of the replication required for reliable assessment. Empirically, overall performance was expected to reach an effective repeatability of 0.80 after approximately 12–22 contests, depending on the group. In the group-specific simulations, the corresponding requirement for at least 90% of populations to recover underlying overall performance at r ≥ 0.80 ranged from 11–21 contests at ρ = 0.3. Under the reference simulation parameterisation, 15 contests were required. These analyses use different definitions of reliability and should not be interpreted as mathematically equivalent, but their convergence indicates that approximately 10–20 or more repeated contests may be required to estimate individual 1v1 performance reliably, with substantially greater replication required when players are relatively similar.

Discussion

Game-based assessments can provide unusually direct measures of competitive performance because the outcome itself may be objectively defined: an attacker succeeds or does not, a defender prevents progression or does not, and scoring outcomes can be recorded without requiring an observer to judge

how well the athlete performed. The principal measurement problem is therefore not necessarily ambiguity about what happened, but attribution: how much does an objectively observed competitive outcome tell us about the individual who produced it? Any single outcome reflects the focal athlete, the opponent encountered and chance variation in how a small number of competitive opportunities unfold. Our results show that this attribution problem can be addressed through deliberate repeated sampling. Single 1v1 contests provided relatively imprecise estimates of attacking, defensive and overall performance, but reliability increased substantially when contests were aggregated. Opponent-adjusted ratings were also stable when estimated independently from repeated subsets of contests, and older cohorts consistently outperformed adjacent younger cohorts in attacking and defensive performance. In the simulations, performance estimated from a limited number of contests increasingly approached each player's expected performance across the population of possible opponents as observations accumulated. Critically, substantially more contests were required when differences among players were smaller. Objective competitive outcomes can therefore provide informative measures of individuals, but only when the competitive contexts generating those outcomes are sampled sufficiently.

This distinction changes how low reliability from a single competitive contest should be interpreted. In a highly standardised individual test, unexplained variation may primarily be regarded as measurement error to be minimised. In a competitive task, some variation arises precisely because the assessment retains the interaction it is intended to measure. An attacker may perform poorly because they encounter a strong defender or because only a few probabilistic opportunities happen to be unsuccessful on that occasion. Greater standardisation could reduce this variation but may simultaneously remove the competitive interaction that makes the outcome representative of performance [4]. The alternative is to sample the sources of attribution uncertainty rather than eliminate them, consistent with the broader treatment of performance assessment as a sampling problem in which observed performance depends on the conditions under which it is elicited [17, 32]. Repeatedly observing players against different opponents reduces the influence of particular matchups and individual contest outcomes, allowing performance consistently associated with the focal player to become more apparent. Our opponent-adjusted analyses illustrate the same principle from another direction: raw and adjusted overall performance were strongly associated, but accounting for the opponents encountered still changed individual rankings by as many as 13 positions. Controlled tests and repeated competitive assessments therefore provide complementary information: the former isolate particular capabilities, whereas the latter estimate competitive success across sampled interactions.

Repeated encounters between the same two players further illustrate why objective outcomes should not be confused with precise individual estimates. Whether a player scored in a particular attacking opportunity was unambiguous, yet the outcomes of repeated contests involving exactly the same players were only moderately consistent. There was no systematic improvement or decline at the second encounter and no consistent evidence that particular attacker–defender pairings produced persistent outcomes beyond the general abilities of the two players. Thus, much of the variation between repeated contests appears to arise because attacking and defensive characteristics determine probabilities of success rather than fixing individual outcomes. Three attacks can therefore produce

different contest scores even when the competitors themselves have not changed. Replication is required not because the competitive outcome is measured poorly, but because any finite sample of competitive outcomes provides an imperfect estimate of the individual.

The age-group analyses provide an independent empirical demonstration that meaningful individual differences nevertheless emerge from these outcomes. Older cohorts had higher attacking and defensive performance in all three adjacent-age comparisons, with attacking odds 1.69–1.83 times greater and defensive advantages 1.49–2.66 times greater than the corresponding younger cohorts. This is consistent with evidence that representative small-sided assessments can discriminate among players differing in age or playing level [14, 23]. The result is important because the individual differences identified by repeated 1v1 assessment corresponded to an independently defined developmental difference rather than simply producing internally repeatable rankings. Attack and defence also provided partly different information, reinforcing the value of retaining these components alongside an overall competitive-performance measure when the task permits.

How much replication is enough?

The amount of sampling required to resolve the attribution problem cannot be specified independently of the population being assessed. Empirically, overall performance was expected to reach an effective repeatability of 0.80 after approximately 12–22 contests across the five groups, while independent split-half analyses produced full-sample reliability estimates of approximately 0.70–0.81 at the observed contest volumes. The group-specific simulations produced a similar pattern: under the intermediate attack–defence correlation of $\rho = 0.3$, at least 90% of simulated populations recovered expected overall performance at $r \geq 0.80$ after 11–21 contests, depending on the empirical population. The empirical and simulation analyses now define overall performance on the same substantive scale—expected net competitive outcome across an opponent population—but quantify different measurement properties. The empirical analyses estimate the repeatability and stability of performance from observed contests, whereas the simulations quantify recovery of known model-specified expected performance. Their numerical thresholds should therefore not be interpreted as mathematically equivalent. Nevertheless, both approaches show that the same competitive task can require different amounts of sampling in different populations.

The controlled simulations explain why. When attacking and defensive between-player variances were doubled relative to the reference condition, the ≥ 90% recovery criterion was reached after nine contests at $\rho = 0.3$. Under reference heterogeneity, 15 contests were required; when the variances were halved, the requirement increased to 27 contests. The same ordering occurred at every assumed attack–defence correlation. Absolute requirements changed with ρ—for reference heterogeneity, from 19 contests at $\rho = 0$ to 15 at $\rho = 0.3$ and 13 at $\rho = 0.6$—but the central heterogeneity effect remained. Populations containing smaller differences among players consistently required more observations to recover those differences with the same certainty. This follows the broader principle that reliability depends on stable between-

individual variation relative to other sources of variation and therefore belongs to a measurement procedure applied to a particular population rather than to a test in isolation [16, 28–30].

This population dependence has an important implication for talent identification. Talent-identification systems commonly reduce large candidate populations through successive selection stages [1, 22, 31]. If selection is effective, later stages should contain fewer but increasingly similar athletes. Fewer athletes reduce the total number requiring assessment, but smaller between-athlete differences increase the information required per athlete. Later-stage assessment may consequently require more observations per athlete, not fewer. Relatively rapid competitive assessment may be sufficient when the aim is to distinguish large differences during broad screening, whereas intensive repeated sampling becomes increasingly important when consequential decisions involve closely matched candidates. Our heterogeneity treatments should not be interpreted as literal representations of particular selection stages, but they demonstrate the statistical principle underlying this prediction.

Designing repeated 1v1 assessments

The present results suggest several practical principles for using objective competitive outcomes to assess individuals. First, the series of contests, rather than an individual contest, should be considered the assessment. Second, opponents should be deliberately varied. The aim is not merely to accumulate more outcomes, but to sample the competitive context against which those outcomes are generated. More generally, performance-assessment theory emphasises that observations sampled from a narrow set of conditions need not generalise to the broader set of conditions about which inference is intended [17, 32]. Ten contests against one opponent may therefore provide different information about an athlete from ten contests distributed across opponents of differing ability. Third, opponent quality should be accounted for analytically when competitive schedules are unequal; otherwise, differences in accumulated success may partly reflect who happened to play whom. Fourth, replication should not simply be copied from another study. Reliability depends on the population as well as the task [16, 28–30], so between-athlete and contest-to-contest variation should be quantified in the population being assessed. Fifth, the required precision should reflect the purpose of the assessment: broad screening or developmental feedback may tolerate greater uncertainty than consequential decisions among closely matched athletes. Finally, distinct competitive roles such as attacking and defending should be retained where meaningful because an overall outcome can conceal different ways in which competitive success is generated.

The approximately 12–22 contests required empirically to reach an effective repeatability of 0.80 should therefore be treated as benchmarks for this protocol and these populations, not universal prescriptions. The simulations reinforce this caution: even with the contest structure held constant, changing only between-player heterogeneity substantially altered the required replication, and uncertainty about the association between attacking and defensive ability changed the absolute thresholds further. Different numbers of competitive opportunities, scoring rules, opponent schedules and population structures should similarly alter how rapidly individual differences emerge. The transferable contribution is

therefore not a recommended number of contests, but a procedure: define an objective competitive outcome, identify the contextual sources affecting its attribution to the individual, sample those sources deliberately, and determine empirically when sufficient information has accumulated for the intended decision.

From 1v1 contests to general principles for game-based assessment

One-versus-one contests provide the simplest implementation of this framework because the attribution problem contains only a focal athlete, an opponent and stochastic variation among competitive opportunities. More complex games add teammates, additional opponents and combinations among players. Repeated 3v3 observations, for example, show that objective game outcomes contain contributions from both individual players and the particular combinations of teammates alongside whom they perform [14, 33]. An objective outcome may still be recorded perfectly—for example, whether a team scores, concedes or retains possession—but attributing that outcome to a particular individual becomes progressively more difficult as additional athletes contribute to it. This problem is also familiar from team-sport analytics, where objectively observed scoring outcomes must be disentangled from the teammates and opponents present [15]. Increasing interpersonal complexity therefore expands the dimensions that must be sampled before competitive outcomes can be attributed reliably to individuals.

Extending the present approach consequently requires identifying what must be replicated, rather than simply asking how many games are needed. In a 1v1 assessment, replication involves additional contests and opponents. In a 2v2 or 3v3 assessment, it may additionally require sampling teammates, opposing combinations and repeated configurations because both individual-player and teammate-combination effects can contribute to observed competitive outcomes [14, 33]. Ten contests against ten different opponents may therefore provide different information from ten contests against one opponent, just as ten small-sided games with systematically rotated teammates may provide different information from ten games within a fixed team. Empirical designs can estimate how much variation is associated with focal athletes, opponents, teammates and their combinations; these estimates can then parameterise simulations that determine how extensively each dimension must be sampled to recover individual performance with the required precision. This shifts the question from simply “how many games?” to “how many games, against how many opponents, with how many teammates, and across how many independent combinations?” The present 1v1 system represents the simplest case of this broader approach and provides a foundation for determining how sampling requirements increase as the competitive network becomes more complex.

Several limitations define the scope of these conclusions. Our empirical analyses involved youth football players and a specific 1v1 protocol containing only three attacking and three defensive opportunities per contest. Replication requirements are therefore expected to change with task structure, scoring rules, opponent scheduling and population characteristics. The simulations also required assumptions about the within-player association between attacking and defensive abilities because their covariance was not

estimated by the empirical crossed models. Sensitivity analyses showed that this assumption affected absolute replication thresholds, although the effect of heterogeneity remained consistent across ρ = 0, 0.3 and 0.6. The controlled heterogeneity conditions were statistical manipulations around an empirically derived reference structure and should not be interpreted as literal populations or particular stages of talent identification. Similarly, age-group differences demonstrate developmental discrimination but are not equivalent to talent or future potential. Finally, the relative simplicity of 1v1 contests means that the present results cannot establish how much sampling will be required when objective competitive outcomes depend simultaneously on multiple teammates, opponents and player combinations. Determining those requirements is an important next step.

In conclusion, objective competitive outcomes solve only the first part of the individual-assessment problem. They provide an unambiguous record of what happened in a competitive interaction, but they do not by themselves establish how much of that outcome should be attributed to the individual athlete. That requires sampling. Across five youth-football populations, repeated 1v1 contests progressively produced more reliable estimates of attacking, defensive and overall competitive performance and consistently distinguished adjacent developmental cohorts. Simulations showed why the required sampling cannot be universal: when differences among players were reduced while the assessment itself was held constant, substantially more contests were required to recover expected individual performance with the same certainty. The general principle is therefore objective outcome → attribution → sampling: define competitive success clearly, identify the other athletes and stochastic processes that influence its attribution, and sample those sources sufficiently for consistent individual differences to emerge. This provides a basis for extending outcome-based assessment from simple 1v1 contests towards more complex games in which teammates, opponents and their combinations must themselves become dimensions of the assessment design.

Abbreviations

1v1: One-versus-one
2v2: Two-versus-two
3v3: Three-versus-three
AIC: Akaike information criterion
AUC: Area under the receiver operating characteristic curve
CI: Confidence interval
GLMM: Generalised linear mixed model
ICC: Intraclass correlation coefficient
OR: Odds ratio
SD: Standard deviation
AI: Artificial intelligence

Declarations

Ethics approval and consent to participate

Ethical approval for this study was granted by The University of Queensland Human Research Ethics Committee (2021/HE001974). Written informed consent was obtained from all players and their parent or legal guardian.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Authors' contributions

R.S.W. contributed to study conception and design, data collection, statistical analysis, simulation development and analysis, interpretation of results, and manuscript drafting, review and editing. A.H.H. contributed to study conception and design, data collection, and manuscript drafting, review and editing. N.S. contributed to study conception and design, data collection, statistical analysis, simulation development and analysis, interpretation of results, and manuscript review and editing. T.C. contributed to study conception and design, data collection, interpretation of results, and manuscript drafting, review and editing. M.S.C. contributed to study conception and design, data collection, statistical analysis, simulation development and analysis, interpretation of results, and manuscript drafting, review and editing. All authors read and approved the final manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We thank all coaches, players and volunteers involved in this study, particularly the Technical Directors for their logistical support during data collection.

Use of generative artificial intelligence

Generative artificial intelligence (AI) tools were used to assist with language editing, manuscript refinement, organisation and improvement of clarity and presentation. All scientific content, study

design, analyses, interpretations and conclusions were developed and verified by the authors. All AI-assisted outputs were critically reviewed and edited by the authors, who take full responsibility for the accuracy, integrity and originality of the final manuscript.

References

1. Johnston K, Wattie N, Schorer J, Baker J. Talent identification in sport: a systematic review. Sports Med. 2018;48(1):97–109. https://doi.org/10.1007/s40279-017-0803-2.
2. Koopmann T, Faber I, Baker J, Schorer J. Assessing technical skills in talented youth athletes: a systematic review. Sports Med. 2020;50(9):1593–611. https://doi.org/10.1007/s40279-020-01299-4.
3. Barraclough S, Till K, Kerr A, Emmonds S. Methodological approaches to talent identification in team sports: a narrative review. Sports. 2022;10(6):81. https://doi.org/10.3390/sports10060081.
4. Pinder RA, Davids K, Renshaw I, Araújo D. Representative learning design and functionality of research and practice in sport. J Sport Exerc Psychol. 2011;33(1):146–55. https://doi.org/10.1123/jsep.33.1.146.
5. Sarmento H, Clemente FM, Afonso J, Araújo D, Fachada M, Nobre P, Davids K. Match analysis in team ball sports: an umbrella review of systematic reviews and meta-analyses. Sports Med Open. 2022;8:66. https://doi.org/10.1186/s40798-022-00454-7.
6. Kempton T, Sullivan C, Bilsborough JC, Cordy J, Coutts AJ. Match-to-match variation in physical activity and technical skill measures in professional Australian Football. J Sci Med Sport. 2015;18(1):109–13. https://doi.org/10.1016/j.jsams.2013.12.006.
7. Mateus N, Gonçalves B, Abade E, Liu H, Torres-Ronda L, Leite N, Sampaio J. Game-to-game variability of technical and physical performance in NBA players. Int J Perform Anal Sport. 2015;15(3):764–76. https://doi.org/10.1080/24748668.2015.11868829.
8. McLaren SJ, Weston M, Smith A, Cramb R, Portas MD. Variability of physical performance and player match loads in professional rugby union. J Sci Med Sport. 2016;19(6):493–7. https://doi.org/10.1016/j.jsams.2015.05.010.
9. Oslin JL, Mitchell SA, Griffin LL. The Game Performance Assessment Instrument (GPAI): development and preliminary validation. J Teach Phys Educ. 1998;17(2):231–43. https://doi.org/10.1123/jtpe.17.2.231.
10. González-Víllora S, Serra-Olivares J, Pastor-Vicedo JC, da Costa IT. Review of the tactical evaluation tools for youth players, assessing the tactics in team sports: football. SpringerPlus. 2015;4:663. https://doi.org/10.1186/s40064-015-1462-0.
11. Kinnerk P, Harvey S, MacDonncha C, Lyons M. A review of the game-based approaches to coaching literature in competitive team sport settings. Quest. 2018;70(4):401–18. https://doi.org/10.1080/00336297.2018.1439390.

12. Fernández-Espínola C, Abad Robles MT, Giménez Fuentes-Guerra FJ. Small-sided games as a methodological resource for team sports teaching: a systematic review. Int J Environ Res Public Health. 2020;17(6):1884. https://doi.org/10.3390/ijerph17061884.
13. Fenner JSJ, Iga J, Unnithan V. The evaluation of small-sided games as a talent identification tool in highly trained prepubertal soccer players. J Sports Sci. 2016;34(20):1983–90. https://doi.org/10.1080/02640414.2016.1149602.
14. Wilson RS, Hunter AH, Camata TV, Foster CSP, Sparkes GR, Moura FA, Santiago PRP, Smith NMA. Simple and reliable protocol for identifying talented junior players in team sports using small-sided games. Scand J Med Sci Sports. 2021;31(8):1647–56. https://doi.org/10.1111/sms.13969.
15. Ghimire S, Ehrlich JA, Sanders SD. Measuring individual worker output in a complementary team setting: does regularized adjusted plus minus isolate individual NBA player contributions? PLoS ONE. 2020;15(8):e0237920. https://doi.org/10.1371/journal.pone.0237920.
16. Brennan RL, Johnson EG. Generalizability of performance assessments. Educ Meas Issues Pract. 1995;14(4):9–12. https://doi.org/10.1111/j.1745-3992.1995.tb00882.x.
17. Brennan RL. Performance assessments from the perspective of generalizability theory. Appl Psychol Meas. 2000;24(4):339–53. https://doi.org/10.1177/01466210022031796.
18. Bredt SGT, Morales JCP, Andrade AGP, Torres JO, Peixoto GH, Greco PJ, Praça GM, Chagas MH. Space creation dynamics in basketball small-sided games. Percept Mot Skills. 2018;125(1):162–76. https://doi.org/10.1177/0031512517725445.
19. Praça GM, Abreu CO, Rochael M, Moreira PD. How reliable are the tactical measures obtained in soccer small-sided games? A test-retest analysis of observational instruments and GPS-based variables. Proc Inst Mech Eng P J Sports Eng Technol. 2025;239(2):216–25. https://doi.org/10.1177/17543371221113925.
20. Hunter AH, Smith NMA, Bitugu BB, Luguterah AW, Wilson RS. Scanning when passing: a reliable and valid standardized soccer test. Biomechanics. 2025;5(3):61. https://doi.org/10.3390/biomechanics5030061.
21. Wilson RS, David GK, Murphy SC, Angilletta MJ Jr, Niehaus AC, Hunter AH, Smith MD. Skill not athleticism predicts individual variation in match performance of soccer players. Proc R Soc B Biol Sci. 2017;284(1868):20170953. https://doi.org/10.1098/rspb.2017.0953.
22. Sarmento H, Anguera MT, Pereira A, Araújo D. Talent identification and development in male football: a systematic review. Sports Med. 2018;48(4):907–31. https://doi.org/10.1007/s40279-017-0851-7.
23. Klingner FC, Huijgen BCH, den Hartigh RJR, Kempe M. Technical–tactical skill assessments in small-sided soccer games: a scoping review. Int J Sports Sci Coach. 2022;17(4):885–902. https://doi.org/10.1177/17479541211049532.
24. Esteves PT. Posture of the opponent and angular relations to the basket constrain decision-making behaviours in 1v1 of basketball. Cuad Psicol Deporte. 2012;12(Suppl 1):17–22.
25. Fujii K, Yoshioka S, Isaka T, Kouzaki M. The preparatory state of ground reaction forces in defending against a dribbler in a basketball 1-on-1 dribble subphase. Sports Biomech. 2015;14(1):28–44.

https://doi.org/10.1080/14763141.2015.1026931.

26. Wilson RS, Smith NMA, Santiago PRP, Camata T, Ramos SP, Caetano FG, et al. Predicting the defensive performance of individual players in one vs. one soccer games. PLoS ONE. 2018;13(12):e0209822. https://doi.org/10.1371/journal.pone.0209822.
27. Wilson RS, Smith NMA, Ramos SP, Caetano FG, Rinaldo MA, Santiago PRP, Cunha SA, Moura FA. Dribbling speed along curved paths predicts attacking performance in match-realistic one vs. one soccer games. J Sports Sci. 2019;37(9):1072–9. https://doi.org/10.1080/02640414.2018.1544110.
28. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217–38. https://doi.org/10.2165/00007256-199826040-00002.
29. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med. 2000;30(1):1–15. https://doi.org/10.2165/00007256-200030010-00001.
30. Vispoel WP, Morris CA, Kilinc M. Applications of generalizability theory and their relations to classical test theory and structural equation modeling. Psychol Methods. 2018;23(1):1–26. https://doi.org/10.1037/met0000107.
31. Vaeyens R, Lenoir M, Williams AM, Philippaerts RM. Talent identification and development programmes in sport: current models and future directions. Sports Med. 2008;38(9):703–14. https://doi.org/10.2165/00007256-200838090-00001.
32. Shavelson RJ, Baxter GP, Gao X. Sampling variability of performance assessments. J Educ Meas. 1993;30(3):215–32. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x.
33. Salinas S, Yonehara S, Sonera M. Quantifying individual and teammate effects on small-sided football performance through repeated-game observations. Footb Stud. 2026;1:100024. https://doi.org/10.1016/j.footst.2026.100024.

Figures

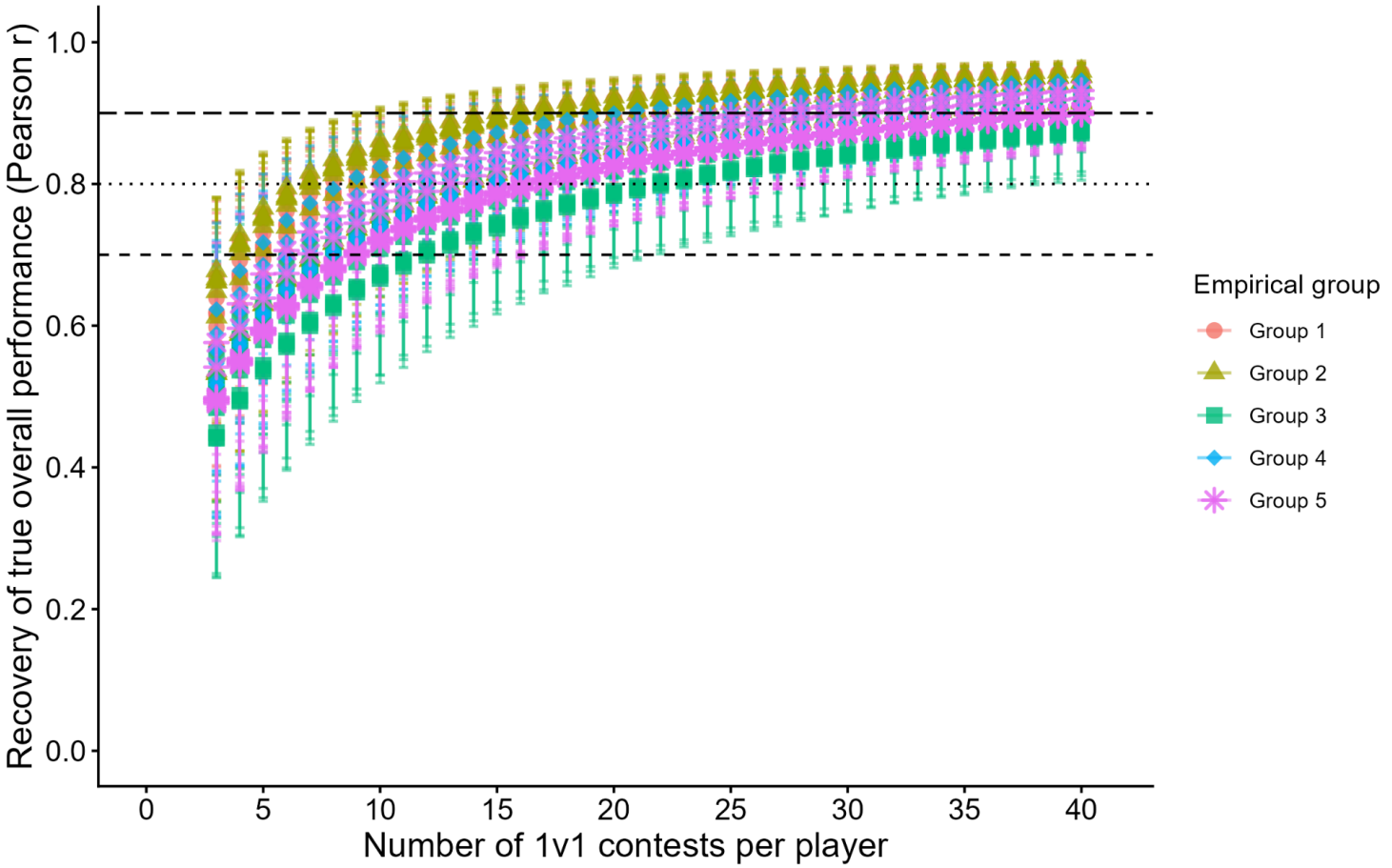

Figure 1

Replication increases the reliability of individual performance estimates from repeated 1v1 contests. Effective repeatability (R_eff) of overall 1v1 performance as a function of the number of contests completed for five empirical player groups. Curves were calculated from the single-contest repeatability (R) estimated independently for each group using the Spearman–Brown relationship. Error bars shown at five-contest intervals represent 95% confidence limits propagated from the 5,000-resample bootstrap confidence intervals for single-contest repeatability. Horizontal reference lines indicate R_eff values of 0.70, 0.80 and 0.90.

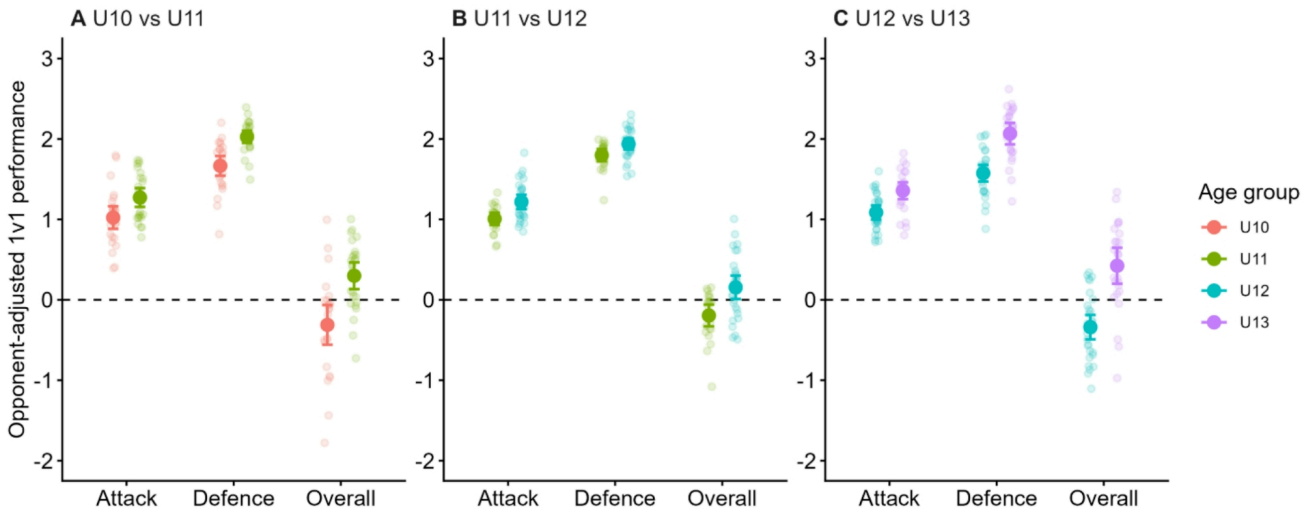

Figure 2

Opponent-adjusted 1v1 performance across adjacent annual age cohorts. Performance is shown for (A) U10 and U11 players in Group 1, (B) U11 and U12 players in Group 3, and (C) U12 and U13 players in Group 2. Small points represent individual player estimates, while larger points and error bars show age-cohort means and 95% confidence intervals. Attacking and defensive performance are model-based expected successes after accounting for opponent identity. Overall performance is the opponent-standardised expected net outcome, calculated as expected goals scored minus expected goals conceded across the relevant opponent population. Formal age comparisons were based on contest-level crossed binomial GLMMs for attacking and defensive success; overall comparisons are descriptive.

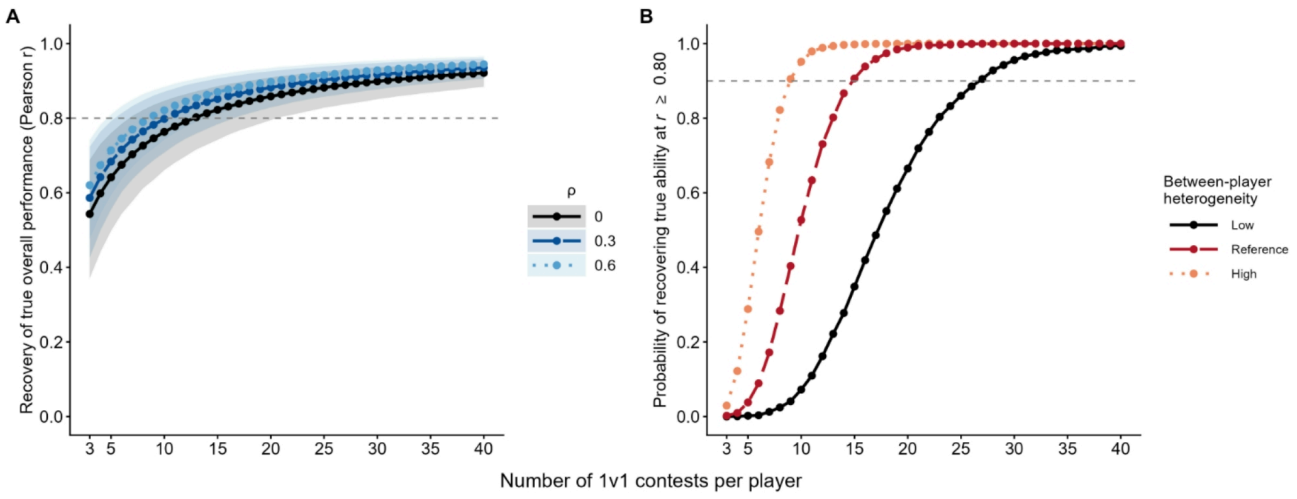

Figure 3

Replication and between-player heterogeneity determine recovery of model-specified overall 1v1 performance. (A) Mean correlation (*r*) between observed overall performance and model-specified expected net performance across 5,000 simulated populations under reference heterogeneity, shown for attack–defence correlations (ρ) of 0, 0.3 and 0.6. Shaded regions represent the 5th–95th percentile range, and the dashed horizontal line indicates $r = 0.80$. (B) Probability that observed overall performance recovered model-specified expected net performance at $r \geq 0.80$ under low (0.5×), reference (1×) and high (2×) between-player heterogeneity at $\rho = 0.3$. The dashed horizontal line indicates a 90% probability of achieving $r \geq 0.80$.